Métodos avançados de previsão:
Análise de fatores de risco clínico e impacto económico numa unidade hospitalar

João Miguel Ramos Chaves Fernandes

Dissertação de Mestrado

Orientador na FEUP: Prof.º António Ernesto da Silva Carvalho Brito

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado Integrado em Engenharia Industrial e Gestão

2015-06-23

Resumo

Na sequência da crise económica e financeira a nível global, desde 2009, os gastos com saúde nos países da OCDE têm vindo a estagnar e, nalguns casos, mesmo a descer. Tal tendência é contrária ao forte investimento crescente na área que se verificava já desde longa data. A crise forçou muitos governos, como é o caso de Portugal, a realizar cortes difíceis em gastos públicos, tendo reduzido os gastos na Saúde em 15% entre 2011 e 2013. Neste contexto, muitos países têm vindo a promover reformas no setor para aumentar a eficiência e produtividade, tentando, idealmente, salvaguardar a qualidade do serviço.

O trabalho aqui apresentado vem de encontro a este problema, sendo o seu objetivo principal o desenvolvimento de uma ferramenta que consiga prever de forma fiável a ocorrência de reinternamentos numa unidade hospitalar.

Para dar resposta ao problema, foram utilizadas técnicas de *Data Mining* e algoritmos preditivos como o *Random Forest* e Redes Neuronais (algoritmo *Multi-layer Perceptron* - MLP), tendo sido estudada a otimização dos parâmetros dos mesmos. Para treino e teste do modelo, foram utilizados dados de um hospital público português, disponibilizados pela Glintt Healthcare Solutions – 36 396 episódios de internamento, ocorridos entre 2012 e 2014.

Nos teste realizados, foi utilizando toda a população, tendo sido realizada uma partição de dados de 70% para treino e 30% para teste, mantendo sempre em cada grupo a proporção original de episódios que resultaram em reinternamentos – 9.25%. Os melhores resultados foram conseguidos com o algoritmo MLP, tendo uma taxa de acerto de 86.36%, com 24.82% de precisão e 23.37% de sensibilidade.

Para além do modelo geral (aplicável a qualquer episódio), testou-se ainda, como prova de conceito, um modelo customizado para prever apenas casos diagnosticados com Pneumonia. Seguindo a mesma forma de teste, conseguiram-se resultados superiores, também usando o algoritmo MLP, obtendo uma taxa de acerto de 77.10%, com 37.08% de precisão e 37.08% de sensibilidade.

A utilização de modelos preditivos tem vindo a tornar-se, cada vez mais, uma ferramenta indispensável para os decisores. O recurso a esta técnica permite reduzir a despesa das instituições hospitalares através do aumento dos níveis de eficiência. Desta forma, é possível assegurar a universalidade da prestação de cuidados de saúde e, simultaneamente, conseguir manter ou até melhorar a qualidade de cuidados prestados aos utentes.

Advanced prediction methods:

Analysis of clinical risk factors and economic impact in a hospital unit

Abstract

Following the recent economic and financial crisis, since 2009, the healthcare expenditure in ODCE countries has been stagnating, and, in some cases, even reducing. This tendency is contrary to the steady increase in investment which had been trending throughout the previous years. The crisis has forced many governments to promote challenging cuts in public expenditure, as is the case for Portugal, which cut 15% on health costs between 2011 and 2013. In this context, many countries have promoted reforms to the health sector in order to increase efficiency and productivity, whilst ideally maintaining the quality of the services.

The work presented here aims to address these issues, as its main goal is to develop a tool which can accurately predict the occurrence of readmissions in a public hospital.

In order to develop this tool, Data Mining techniques were employed, as well as predictive algorithms, such as Random Forest and Neural Networks, specifically the Multi-Layer Perceptron algorithm. The optimization of the parameters for the aforementioned algorithms was also studied. For training and testing the model, data from a Portuguese public hospital was employed, made available by Glintt Healthcare Solutions – 36 396 inpatient episodes, occurred between 2012 and 2014.

The entire population was employed in each of the tests conducted: a partition of 70% of the data was used for training and 30% for testing, while maintaining, in each group, the original proportion of readmission cases – 9.25%. The MLP algorithm achieved the best results, with an Accuracy of 86.36%, a Precision of 24.82% and Recall of 23.37%.

Besides this general model, applicable to any inpatient episode, as a proof of concept, an additional model was developed, customized to predict readmissions regarding cases diagnosed with Pneumonia. Using the same testing method as the general model, superior results were achieved, using the MLP algorithm, resulting in a model with an Accuracy of 77.10%, Precision of 37.08% and Recall of 37.08%.

The use of predictive models is increasingly becoming an essential tool for decision-makers. Resorting to these techniques allows for a reduction in hospital expenditure through increasing efficiency. This way, it is possible to ensure universal access to healthcare, while simultaneously maintaining, or even improving the quality of the services.

Agradecimentos

Gostaria de agradecer, nesta secção, a todas as pessoas que contribuíram para a realização desta tese.

Em primeiro lugar, gostaria de agradecer ao professor António Carvalho Brito, pela sua orientação ao longo da tese e ajuda na estruturação do presente documento. Agradeço também ao professor José Luís Borges pela ajuda na compreensão de alguns conceitos de *Data Mining* e ao professor João Mendes Moreira pela disponibilização de conteúdos e bibliografia para estudo da mesma área.

Agradeço também à equipa do SIG pela disponibilização do servidor, dos dados analisados, ajuda na compreensão dos mesmos, resolução de problemas que tivessem e todo o tempo e recursos envolvidos, nomeadamente ao Pedro, Paulo, Hugo e ao Maurício. Agradeço ainda ao Rui Gouveia pela orientação relativamente ao tratamento dos diagnósticos e registo da informação.

Deixo um enorme agradecimento ao Steeve, meu orientador na Glintt, cujo apoio foi crucial em todos os pontos do projeto, especialmente na construção do modelo, bem com à Ilda, que tanto me ajudou na compreensão do sistema de saúde e na revisão do documento. Agradeço também ao José pela ajuda na programação em R e na elaboração de gráficos e ao Ricardo pela ajuda na elaboração de apresentações.

Agradeço também ao professor Nuno Ramos, meu tio, pela ajuda na estruturação da tese e à minha prima Teresa pelo conhecimento e experiência partilhada sobre a área da saúde.

Agradeço à minha família, especialmente aos meus pais e ao meu irmão por todo o apoio ao longo do curso e me terem proporcionado todas as condições para chegar até aqui.

Agradeço, por fim, à Glintt, pela oportunidade de realização de um projeto tão aliciante e por me acolher ao longo do seu desenvolvimento.

Índice de Conteúdos

Siglas

ACSS	Administração Central do Sistema de Saúde
ARS	Administração Regional de Saúde
BI	*Business Intelligence*
CCS	*Clinical Classifications Software*
DDR	Dose Diária Recomendada
DPCO	Doença Pulmonar Obstrutiva Crónica
DRG	*Diagnosis Related Groups*
FN	Falsos Negativos
FP	Falsos Positivos
GDH	Grupos de Diagnósticos Homogéneos
HCUP	*Healthcare Cost and Utilization Project*
ICD-9	*International Classification of Diseases*, 9ª revisão
ICM	Índice de *Case-Mix*
KDD	*Knowledge Discovery in Databases*
MLP	*Multi-Layer Perceptron*
NN	*Neural Networks*
OMS	Organização Mundial da Saúde
RF	*Random Forest*
SIDA	Síndrome de Imunodeficiência Adquirida
SIG	Sistema de Informação de Gestão
SNS	Sistema Nacional de Saúde
SVM	*Naive Bayes* e *Support Vector Machines*
UE	União Europeia
ULS	Unidade Local de Saúde
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos

Índice de Figuras

Índice de Tabelas

1 Introdução

Esta dissertação resulta de uma proposta de projeto lançada pela Glintt Healthcare Solutions S.A., enquadrando-se no âmbito do Projeto de Dissertação do Mestrado Integrado em Engenharia Industrial e Gestão da Faculdade de Engenharia da Universidade do Porto (FEUP).

O objetivo último deste Projeto é melhorar a Qualidade dos Serviços prestados pelas instituições de Saúde pertencentes ao Serviço Nacional de Saúde (SNS) e diminuir os encargos financeiros associados a Episódios de Internamento Hospitalar. Pretende-se, nesse sentido, melhorar o processo de atribuição de Altas Hospitalares, bem como de Identificação de Fatores de Risco de Reinternamento.

Tais melhorias são conseguidas através de uma ferramenta de apoio à decisão, que consiga prever de forma fiável se um paciente será ou não reinternado, usando, por base, a Informação Clínica do mesmo. Para o desenvolvimento desta ferramenta, foi utilizada informação real de um Hospital Público que, por motivos de confidencialidade, não poderá ser identificado.

1.1 Apresentação da Empresa

A Glintt – Global Intelligent Technologies é uma das maiores empresas tecnológicas Portuguesas, que opera na Europa, África e América Latina. Tem uma presença forte nos setores da Banca, Telecomunicações, Saúde, Comércio, Indústria e Administração Pública (Técnico 2008).

A Glintt Healthcare Solutions é uma empresa integrada no grupo Glintt, que conta já com mais de 20 anos de experiência no setor da Saúde. Dedica-se à prestação de serviços na área das tecnologias da informação, marcando presença em mais de 200 hospitais e clínicas (Glintt 2013).

A Glintt Healthcare Solutions está dividida em diferentes unidades/equipas funcionais. O projeto foi desenvolvido com o apoio da equipa de Sistema de Informação de Gestão (SIG). O SIG tem como principais objetivos disponibilizar indicadores de produção, qualidade e financeiros, dotando as organizações de saúde com uma ferramenta de *reporting* de gestão. Desta forma, a ferramenta desenvolvida e disponibilizada pelo SIG permite às instituições de saúde avaliar e acompanhar (em tempo real) indicadores de acesso, de desempenho assistencial e económico-financeiros (Gama et al. 2012).

1.2 O Projeto "Previsão de Risco Clínico" e Objetivos

Casos de reinternamento implicam custos significativos tanto para os pacientes como para as organizações de saúde (Sousa-Pinto et al. 2013). A taxa de reinternamentos hospitalares é usada como um dos indicadores de desempenho dos hospitais, sendo um dos fatores utilizados para definir a atribuição de fundos e incentivos (ou penalizações) monetários por parte do SNS (ACSS (Ministério da Saúde) 2015).

Uma das soluções desenvolvidas pela Glintt é um Sistema de Gestão de Informação Hospitalar, que faz uso de diversas técnicas de *Business Intelligence* (BI), disponibilizando um *report* de gestão às instituições de saúde. No entanto, tal ferramenta, não possui ainda uma componente preditiva. Nesse sentido, através deste Projeto, um dos objetivos da Glintt será adicionar as funcionalidades da ferramenta à solução referida.

O projeto terá, como objetivo principal, desenvolver um algoritmo que consiga, com base na informação clínica armazenada em *Data Warehouse*, identificar corretamente se um episódio irá resultar num Reinternamento.

Para além disso, pretende-se também identificar os fatores (tais como a idade, género ou patologia) que mais influência têm sobre os reinternamentos. Finalmente, um outro objetivo será estimar os ganhos monetários e de eficiência que a implementação da ferramenta traria ao Hospital em análise.

Com vista a dar resposta a este problema, o Projeto foi dividido num conjunto de etapas sequenciais:

- Recolha bibliográfica relativa a diversas áreas:
    - *Knowledge Discovery in Databases* (KDD) – estado da arte de Metodologias e Algoritmos Preditivos;
    - Data Mining – Escolha das Ferramentas mais adequadas a trabalhar os dados;
    - Saúde – Aplicação de técnicas de Data Mining a prolemas semelhantes;
    - Saúde – conhecimento já documentado sobre Reinternamentos Hospitalares;
- Recolha e análise de dados;
- Tratamento de dados;
- Implementação e teste de Algoritmos preditivos;
- Avaliação de Resultados e melhoria Iterativa (nove implementação);
- Desenvolver uma metodologia de estimar os ganhos para o hospital e em qualidade de serviço.

1.3 Método seguido no projeto

O Projeto "Previsão de Risco Clínico" trata-se, fundamentalmente, de um projeto de KDD na área da Saúde.

Foi seguida a Metodologia CRISP-DM, sendo a mais utilizada atualmente em projetos de KDD (Piatetsky 2014), estando já estabelecida como *standard* da indústria (Marbán, Mariscal, and Segovia 2009).

Uma parte fundamental de um projeto de KDD assenta no conhecimento empírico da área que se pretende analisar, pelo que se torna indispensável incorporar o conhecimento de Profissionais de Saúde.

No desenvolvimento do Projeto, foram consultados vários médicos, farmacêuticos, enfermeiros e consultores da área de Saúde por forma a discernir qual seria a informação relevante, bem como qual a melhor forma de a incorporar nos dados.

1.4 Estrutura da dissertação

A presente tese contempla cinco capítulos e diversos anexos.

No Capítulo 2 serão abordados os temas da Saúde em Portugal, o estado da arte em KDD e a sua aplicação à área da Saúde. Neste, irá apresentar-se, de forma sucinta, os órgãos do SNS, bem como o funcionamento da Contratualização dos Serviços de Saúde em Portugal e a atribuição de fundos e incentivos a instituições de saúde. Seguidamente, irá demonstrar-se a relevância em utilizar a taxa de reinternamento como um indicador da qualidade de serviço em hospitais. Após isso, será apresentada a metodologia utilizada e os algoritmos aplicados na abordagem seguida. Finalmente, serão apresentados alguns exemplos da aplicação recente de KDD à área da saúde.

O Capítulo 3 é dedicado a explicar, detalhadamente, os passos seguidos na construção do modelo preditivo. Neste capítulo, será abordada a identificação dos dados relevantes para análise, a construção do *dataset* utilizado no modelo, desde a recolha, tratamento, limpeza e filtragem, bem como alguns campos relevantes calculados e o método de cálculo. Após isso, serão apresentados os vários modelos e algoritmos testados, bem como as variantes do *dataset* utilizados em cada um e, finalmente, o método e métricas utilizadas na avaliação do desempenho dos modelos e algoritmos.

No Capítulo 4 irá apresentar-se os resultados obtidos para cada um dos algoritmos e modelos testados, bem como a interpretação dos mesmos. Para além disso, serão expostas as análises estatísticas elaboradas relativamente aos atributos mais relevantes utilizados no modelo.

Por fim, o Capítulo 5 conta das conclusões do projeto e propostas de trabalhos futuros, tais como melhorias no modelo, dados novos a incorporar e aplicações do mesmo.

2 Aplicabilidade de *Knowledge Discovery in Databases* aos Serviços de Saúde em Portugal

2.1 Contratualização dos Serviços de Saúde

2.1.1 O Sistema Nacional de Saúde

Portugal, como Estado-Membro da União Europeia (UE), orienta-se pelos valores e princípios comuns aos Estados, deliberados em Conselho Europeu em 2006. Neste, os Sistemas de Saúde foram consagrados como parte fundamental no que respeita à proteção social na Europa. A UE pretendeu, desta forma, assegurar o acesso a cuidados médicos a todos os cidadãos europeus, através de um novo programa que conseguisse dar resposta aos desafios económicos e demográficos que a Europa tem recentemente vindo a enfrentar.

O SNS foi criado em 1979, a 15 de Setembro, pela lei nº 56/79. Através dele, o acesso a cuidados médicos passou a ser garantido e gratuito a todos os cidadãos portugueses. No entanto, foram introduzidas as taxas moderadoras em 1992, como forma de racionalizar o consumo e a utilização de cuidados de saúde, não sendo, desta forma, inteiramente gratuito.

O SNS, em termos de gestão, possui autonomia administrativa e financeira, tendo uma estrutura descentralizada. O SNS compreende órgãos centrais, como a Administração Central do Sistema de Saúde (ACSS), regionais, como as Administrações Regionais de Saúde (ARS), e locais, como as Unidades Locais de Saúde (ULS). Dispõe de serviços prestadores de cuidados de saúde primários (Centros de Saúde), de saúde secundários (Centros Hospitalares e Hospitais), de saúde terciários e de outros institutos.

Atualmente, desde a publicação de um Decreto-Lei em 2011, o SNS integra todos os serviços e entidades públicas prestadoras de cuidados de saúde. De entre estes, salienta-se os Agrupamentos de Centros de Saúde, os estabelecimentos hospitalares, independentemente da sua designação, e as unidades locais de saúde (Boquinhas 2012).

2.1.2 Contratualização entre ARS e Hospitais

A prestação de serviços Hospitalares em Portugal, para instituições Públicas, é contratualizada todos os anos entre as Instituições e a ARS da região, ficando registada no documento Contrato-Programa do respetivo ano. A ACSS disponibiliza um documento, de acesso público, no qual se encontram os princípios orientadores desse processo contratual. Neste, encontra-se também o orçamento máximo alocado para cada ARS e ULS, a forma de cálculo da contratualização ARS-Hospital bem como os incentivos ao cumprimento de determinadas metas de indicadores de qualidade e eficiência de serviço (ACSS (Ministério da Saúde) 2015).

Atualmente, o SNS utiliza um preço base único por episódio Hospitalar para a atividade agrupada em Grupos de Diagnósticos Homogéneos (GDH) e cálculo do índice de *case-mix*. Este índice é calculado de acordo com a produção de internamento e ambulatório relativa ao ano completo mais recente com informação disponível.

Os GDH são um sistema de classificação de doentes internados em hospitais de agudos, correspondendo à tradução portuguesa do sistema de origem americana de classificação *Diagnosis Related Groups* (DRG). Este sistema agrupa-os em grupos clinicamente coerentes e similares do ponto de vista do consumo de recursos (Borges 2011b). Assim sendo, estes permitem definir operacionalmente os produtos de um hospital, ou seja, o conjunto de bens e serviços que cada doente recebe em função das suas necessidades e da patologia. Os GDH descrevem deste modo os gastos médios incorridos para tratamento dos diagnósticos respetivos, sendo usados como base para o financiamento prospetivo (Castro 2011).

Considerando os doentes agrupados por GDH, é calculado todos os anos, com base em dados estatísticos, o custo médio do doente típico desse GDH a nível nacional, que figura na tabela de preços por GDH da ACSS, nos anexos do documento (ACSS (Ministério da Saúde) 2015).

Essa tabela é utilizada para calcular o preço por episódio que cada hospital irá receber no ano seguinte, através do Índice de *Case-Mix* (ICM). Este índice resulta do rácio entre o número de doentes equivalentes ponderados pelos pesos relativos dos respetivos GDH e o número total de doentes equivalentes.

O ICM nacional é, por definição, igual a um. O ICM de cada hospital irá afastar-se mais desse valor de referência, consoante o tratamento de uma proporção maior ou menor de GDH com elevado peso relativo, face ao padrão nacional.

O ICM é um dos fatores utilizados na fórmula de cálculo da remuneração dos hospitais em sede de Contrato-Programa. A fórmula base usada é:

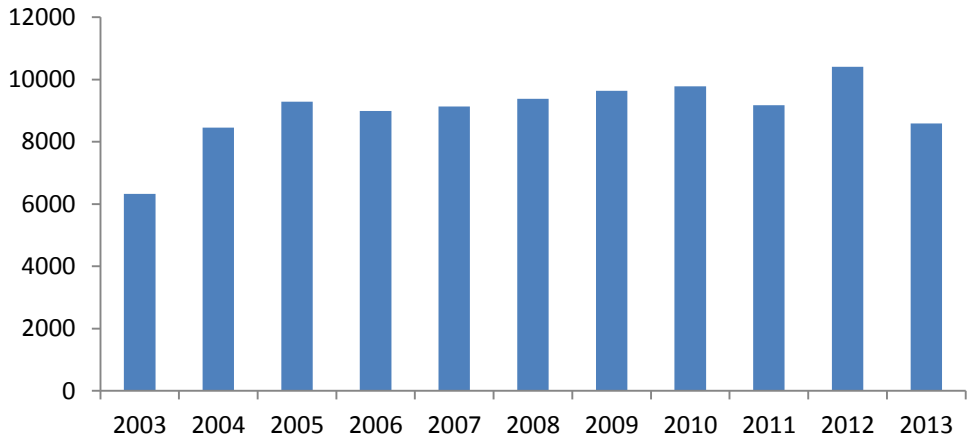
$$Remuneração = Nº\ de\ Doentes\ Equivalentes \times Preço\ Base \times ICM\ Hospital \quad [1]$$

Relativamente à Equação 1, é de referir que o Preço-base é afetado por vários fatores, como a região, a densidade populacional e mesmo pelo Hospital em questão (Borges 2011c). Note-se ainda que, como podem surgir casos de doentes excecionais ou de doentes transferidos, o número de Doentes Equivalentes serve como uma aproximação ao número real de doentes tratados num GDH, permitindo considerar a produção global do internamento como o número de doentes equivalentes e exprimir assim o financiamento desses doentes em termos equivalentes ao financiamento dos denominados episódios normais.

Os episódios são ponderados de acordo com os limites de tempo de internamento considerados normais para o GDH, podendo, por exemplo, uma transferência contar como 0,5 doentes (Borges 2011a). Num episódio típico ou normal um doente saído corresponde a um doente equivalente.

2.2 Taxa de Reinternamento como Indicador da qualidade de serviço

Hoje em dia, no contexto de crise económica que tem vindo a afetar os países ocidentais, é cada vez mais imperativo aumentar a eficiência no setor hospitalar, minimizando os desperdícios. Para tal, torna-se necessário recorrer a indicadores de eficiência e qualidade que sejam relevantes, por forma a medir corretamente o desempenho neste setor e, do ponto de vista governamental, atribuir incentivos monetários.



Fonte:(Fundação Francisco Manuel dos Santos (Instituição) 2015)

Figura 1 - Despesa do Estado em Saúde (milhões de euros)

Já vários estudos evidenciaram que elevadas taxas de reinternamentos estão relacionadas com um aumento das despesas no sector da saúde, correspondendo a uma porção considerável das mesmas (Jweinat 2010). Refletem, para além disso, um decréscimo da qualidade dos cuidados de saúde e estão normalmente associadas a elevadas taxas de mortalidade hospitalar (Sousa-Pinto et al. 2013).

É de notar que, no entanto, as taxas de reinternamentos devem ser medidas relativamente a patologias específicas, sendo falacioso olhar para a taxa geral. Tal deve-se a uma grande percentagem dos reinternamentos (52 a 91%) totais não ser previsível atendendo a que a grande maioria dos episódios trata-se de casos agudos, não se conhecendo, *a priori*, a evolução patológica dos episódios em causa. Por outro lado, para casos de doenças crónicas há uma forte relação causa-efeito entre a vigilância e tratamento médico e a evolução dos sinais e sintomas das destas doenças (Benbassat and Taragin 2000).

Estima-se ainda que entre 12% e 75% dos reinternamentos possam ser prevenidos por melhor educação dos pacientes, melhor avaliação/preparação pré-alta e melhores cuidados domiciliários (Benbassat and Taragin 2000).

Um reinternamento define-se como um internamento do doente no mesmo hospital, num período de setenta e duas horas a contar da data da alta. Da definição excetuam-se:

- Quando o episódio de internamento subsequente não está clinicamente relacionado com o anterior, e as situações do foro oncológico;
- Quando, no internamento inicial, o doente tem saída contra o parecer médico;
- Quando o doente é transferido para realização de exame que obrigue a internamento, seguindo-se o tratamento no hospital de origem (Silva 2011).

Atualmente são já atribuídos incentivos com base em indicadores como a taxa de reinternamento, que constituem 5% do valor dos Contratos-Programa. Ao cumprimento de uma taxa de reinternamentos máxima, encontra-se indexado 4% do total dos incentivos atribuídos (ACSS (Ministério da Saúde) 2015).

Para além destes incentivos, os hospitais têm ainda penalizações associadas à taxa de reinternamento, que variam de instituição para instituição. No do presente Projeto, a instituição em análise não é remunerada pelo SNS por qualquer episódio que seja um reinternamento. Para além disso, caso não consiga ter uma taxa abaixo de um valor estipulado no presente ano, irá sofrer uma penalização significativa sobre a sua remuneração total.

Por razões de confidencialidade, não poderão ser divulgados valores específicos.

2.3 *Knowledge Discovery in Databases*

Atualmente, com a banalização do acesso à tecnologia e o uso generalizado de dispositivos como *smartphones* e *tablets*, a informação gerada pelos seres humanos atingiu níveis nunca antes imaginados.

A análise destes dados é, hoje em dia, uma parte indispensável do negócio de muitas organizações. Estas quantidades de informação, conhecidas como *Big Data* não são passíveis de ser processadas manualmente, requerendo *software* específico e recurso a técnicas de *Data Mining* para extraír delas informação relevante.

Data Mining é uma área interdisciplinar do conhecimento que incorpora métodos de inteligência artificial, *machine learning*, reconhecimento de padrões e estatística para descobrir padrões em grandes *sets* de dados.

No entanto, *Data Mining* é apenas a parte analítica de um processo maior – KDD. Este processo engloba não só o processamento dos dados e aplicação de algoritmos, mas também os processos anteriores (compreensão da área em análise e extração de dados) e posteriores (teste e validação de modelos e implementação da solução), (Marbán, Mariscal, and Segovia 2009).

Os modelos resultantes deste processo inserem-se geralmente em duas categorias principais: modelos Preditivos (supervisionados) e Descritivos (não-supervisionados).

Os modelos Preditivos têm por objetivo, com base num conjunto de dados de treino que tenham valor ou classificação conhecida, atribuir um valor/classificação, que caracterize um novo exemplo. Chamam-se de supervisionados por ser conhecido já, à partida, quais as classificações possíveis para os grupos de dados, ao invés de ser necessário identificar ou definir grupos de entre os dados.

Os modelos Descritivos têm como objetivo descobrir padrões interpretáveis pelo ser humano, como por exemplo agrupar dados em conjuntos com características semelhantes.

Estas duas categorias não são mutuamente exclusivas, havendo modelos que se enquadram em ambas. Estas categorias podem ainda ser subdivididas:

Os algoritmos de modelos Descritivos podem ser de:

- Agrupamento - agregam os dados de acordo com a sua semelhança;
- Associação - procuram encontrar padrões de associação frequentes entre os atributos de um conjunto de dados;
- Sumarização - procuram encontrar uma descrição simples e compacto para grupos de dados.

As tarefas preditivas podem ser subdivididas em duas categorias: Classificação e Regressão. A diferença entre elas é o tipo de dados que classificam:

- Algoritmos de Classificação tratam de previsões discretas, prevendo classes já conhecidas à partida, como uma classificação binária Sim/Não.

- Algoritmos de Regressão, por sua vez, tentam prever valores de dados contínuos, tal como o preço de ações.

O projeto "Previsão de Risco Clínico" baseia-se, assim, na aplicação de Métodos Preditivos de Classificação à informação armazenada dos pacientes por forma a obter um classificador que consiga prever corretamente casos de reinternamento.

2.3.1 Metodologia CRISP-DM

A Metodologia *Cross Industry Standard Process for Data Mining*, ou CRISP-DM divide o processo de KDD em seis fases principais.

A sequência entre as fases não é rígida, sendo que regressar a fases anteriores faz parte natural do processo, dada a sua natureza iterativa. No diagrama (Figura 2) seguinte encontram-se indicadas as fases do processo e as dependências mais importantes e frequentes entre as mesmas.

A seta exterior no diagrama simboliza a natureza cíclica do processo de KDD. É de realçar que este continua mesmo após a implementação da solução.

A aprendizagem ao longo destas etapas despoleta, frequentemente, novas questões de negócio relevantes mais focadas e o processo de KDD subsequente irá beneficiar desta experiência passada. Desta forma, o processo torna-se num ciclo de melhoria iterativa, obtendo uma solução cada vez mais ajustada à realidade que se pretende modelar.

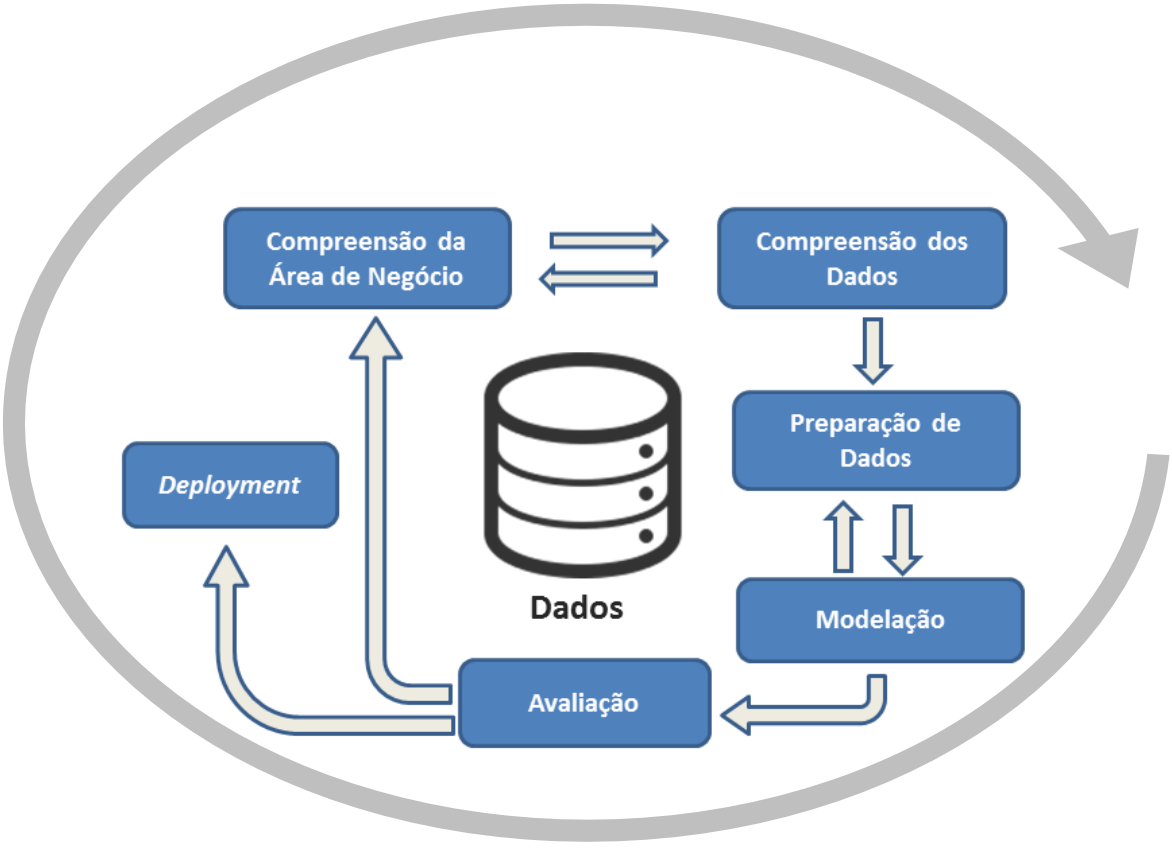


Figura 2 - Fases do processo de KDD

Apresentam-se, de seguida, as seis fases do processo de KDD (Marbán, Mariscal, and Segovia 2009) e o que cada uma envolve:

1. **Compreensão da área de negócio:**
    - Ponto de partida para qualquer projeto de KDD;
    - Compreender os objetivos do projeto e os requisites do ponto de vista do negócio;
    - Converter esse conhecimento num problema de *Data Mining*.
2. **Compreensão dos dados:**
    - Recolha dos dados – obtenção dos dados a analisar; pode envolver, por exemplo, realizar extrações de dados de um servidor ou *query*'s a uma base de dados;
    - Análise inicial dos dados, por forma a familiarizar-se com o seu conteúdo, identificar problemas de qualidade (dados errados ou incompletos, ruído) e retirar as primeiras conclusões, tais como identificar subconjuntos relevantes;
    - Formulação de hipóteses relativas a informação oculta nos dados, tais como correlações ou novos campos que ajudem a classificar os dados que necessitem de ser calculados a partir de um ou mais campos existentes.
3. **Preparação dos dados:**
    - Conjunto de todas as atividades para construir o *dataset* final (dados que irão alimentar os algoritmos preditivos);
    - Este conjunto de tarefas acaba por ter que ser realizado múltiplas vezes, com a adição de novos dados ao modelo ou ao realizar alterações ao *dataset* final;
    - Estas tarefas incluem a seleção de tabelas, registos e atibutos, bem como transformações de atributos e limpeza de dados.
4. **Modelação**
    - Aplicação de técnicas de modelação (e.g. algoritmos preditivos ou de *clustering*);
    - Calibração dos parâmetros dos algoritmos para obter soluções otimizadas;
    - Normalmente, há múltiplas técnicas para resolver o mesmo problema, sendo que algumas têm requisitos específicos relativamente à forma dos dados. Tal pode envolver, por exemplo, apenas aceitar *inputs* binários ou ter incapacidade de processar campos de texto ou campos incompletos. Dessa forma, é frequentemente necessário regressar à fase de preparação de dados.
5. **Avaliação**
    - Avaliação dos resultados do modelo através de métricas apropriadas de um ponto de vista de análise de dados;
    - Revisão dos passos executados na construção do modelo e avaliar se cumpre os requisitos da perspetiva do negócio;
    - É importante determinar se houve alguma questão relevante do negócio que não tenha sido tida em consideração;
    - Decisão de como aplicar os resultados do modelo ou se é necessário reformulá-lo, podendo ter que se voltar a uma etapa anterior do processo.

6. ***Deployment***

- A criação do modelo não é o fim do projeto. O conhecimento gerado por este necessita ainda de ser organizado e apresentado de forma ao cliente poder utilizá-lo;
- Essa tarefa pode ir desde um simples relatório até à implementação de um sistema que seja repetível periodicamente ou implementado numa solução mais abrangente.

2.4 Algoritmos de classificação supervisionados

No caso em análise irão utilizar-se algoritmos preditivos, uma vez que as classes em que se pretende classificar os episódios são conhecidas – gerou/não gerou reinternamento.

Os algoritmos testados foram os seguintes:

- *Random Forest* (RF) - baseado na combinação de árvores de decisão;
- *Neural Networks* (NN) - baseado no funcionamento de neurónios.

Estes algoritmos conseguem, a partir de um número elevado de exemplos, deduzir as relações entre as variáveis dos dados do paciente e criar, com base nelas, um classificador que, face a um novo episódio, retorne uma previsão sobre este resultar (ou não) num reinternamento.

A escolha destes algoritmos foi realizada com base nos resultados de uma tese realizada sobre um *dataset* semelhante àquele em análise, realizada por Gonçalo Pereira (Pereira 2014), na mesma empresa, em que este projeto se baseia.

Na análise realizada, concluiu que, dos algoritmos testados, o *Random Forest* conseguia os melhores resultados. Nos trabalhos futuros da tese, sugere ainda o teste do algoritmo de rede neuronal *Multi-Layer Perceptron* (MLP) que, pela dimensão dos dados em análise e as características do algoritmo, poderia conseguir resultados superiores aos já obtidos.

Para além do RF, testou ainda os algoritmos *Naive Bayes* e *Support Vector Machines* (SVM) com que obteve resultados inferiores.

2.4.1 *Random Forest*

Random Forest (Breiman 1999) é um classificador robusto desenvolvido por Leo Breiman e Adele Cutler. É um classificador de *ensemble*, atendendo a que utiliza vários métodos preditivos combinados – neste caso, múltiplas árvores de decisão distintas geradas a partir dos dados de treino.

Cada árvore irá classificá-lo separadamente, sendo a classificação com mais ocorrências considerada como final. Neste classificador será definido um número de árvores a gerar.

Por forma a reduzir a correlação entre as diversas árvores de decisão, pode aplicar-se o método *Random-Input* para a seleção dos atributos de cada árvore:

Para cada árvore da floresta irá ser selecionado um grupo de X atributos escolhidos aleatoriamente do conjunto. Estes atributos serão utilizados nas decisões dessa árvore.

A divisão dos dados segundo cada atributo será testada por forma a identificar aquele que conduz à maior redução da entropia do conjunto após a divisão. Esta divisão será traduzida na separação dos dados em dois "ramos" da árvore de decisão através de pares de condições mutuamente exclusivas do tipo: $x > 5{,}5; x \leq 5{,}5$, sendo *x* o valor do atributo Y.

As divisões serão realizadas iterativamente expandindo a árvore de decisão até não se conseguir obter mais ganhos de entropia.

Entropia é uma medida de desordem em dados. Havendo, para um certo atributo, N valores possíveis, ou seja N classes, cada uma com uma proporção relativa de p_c no conjunto, sendo:

$$p_c = \frac{numero\ de\ elementos\ da\ classe\ c}{numero\ de\ elementos\ da\ populaçao},$$

a entropia *Ent* do conjunto, atendendo aos valores de um atributo, pode calcular-se pela fórmula:

$$Ent = -\sum_{c=1}^{N} p_c . \log_2 p_c$$

A aleatoriedade na escolha de atributos é introduzida no algoritmo de forma a minimizar a correlação entre as árvores, mantendo uma baixa taxa de erro. É importante evitar demasiada correlação, pois isso implica que são geradas árvores muito semelhantes que irão utilizar os mesmos atributos como nós.

O uso de seleção de atributos aleatório permite que o *Random Forest* tenha boa taxa de acerto e que se mantenha robusto e resistente a *outliers* e ruído, tornando-o superior a outros algoritmos semelhantes como o *Adaboost*. Este processo de ramificação irá ser repetido até a árvore gerada não poder crescer mais.

2.4.2 *Neural Networks*

No âmbito de *machine learning* e ciência cognitiva, redes neuronais artificiais são uma família de modelos de aprendizagem estatísticos inspirados pela biologia das redes neuronais reais do sistema nervoso central de animais, em particular, do cérbero.

São usadas para estimar ou aproximar funções desconhecidas com dependência de um grande número de *inputs*. Geralmente, apresentam-se como sistemas de "neurónios" ligados que enviam mensagens entre si. As ligações têm pesos numéricos que podem ser ajustados com base na experiência (dados de treino, neste caso), tornando as redes neuronais (NN) capazes de aprendizagem ao adaptar-se aos *inputs*.

Multi-Layer Percepton (MLP)

No MLP, cada nódulo (neurónio) irá realizar a soma do produto entre os vários *inputs* recebidos e o peso associado à respetiva ligação, como representado na Figura 3. Essa soma irá passar por uma função de ativação não linear, aproximadamente binária, sendo o resultado deste cálculo propagado para a camada seguinte.

Este processo, conhecido como *forward propagation*, será repetido em todas as camadas até chegar à camada de *output*, os nódulos desta camada final encontram-se associados a cada uma das classes presentes no conjunto de dados.

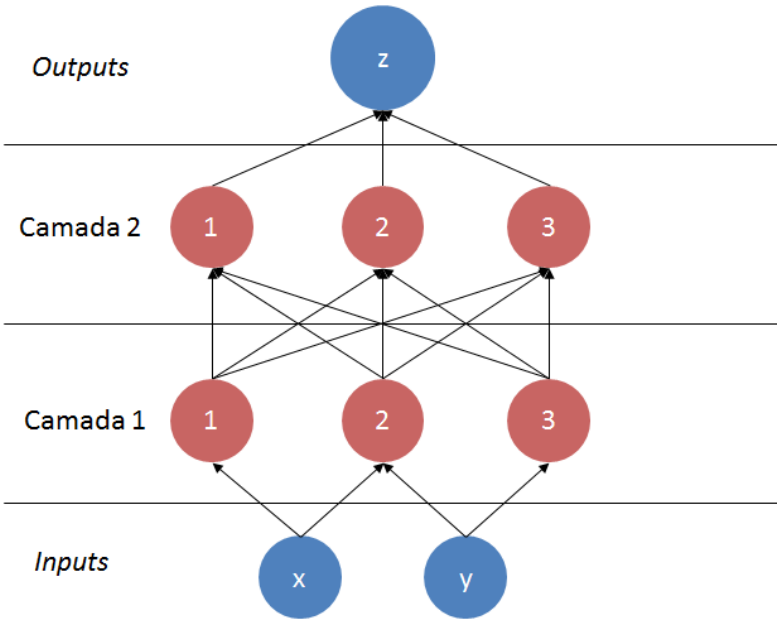


Figura 3 - Diagrama de MLP

Para a realização da aprendizagem, uma das técnicas mais usadas com o MLP é a *resilient backpropagation* (Rprop), uma heurística de otimização de métodos de *backpropagation* (Riedmiller and Braun 1993). O método de *backpropagation* está dividido em dois passos:

O primeiro passo consiste na realização do método de *forward propagation* mencionado e a obtenção do valor de saída, este irá ser comparado com o resultado esperado. A diferença entre estes dois valores será o erro.

No segundo passo o erro será propagado pela rede no sentido inverso de forma aos vários nódulos ajustarem os pesos das ligações com o objetivo de diminuir o erro nas próximas iterações. Através deste processo os MLP realizam a aprendizagem e criam modelos capazes de realizar classificações.

2.4.3 Métricas de Avaliação de Desempenho de Algoritmos

Matriz de Confusão

A matriz de confusão (Figura 4), é uma forma de representação dos resultados de teste de algoritmos preditivos. Para classificações binárias, trata-se de uma matriz de 2x2 onde se registam os Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN).

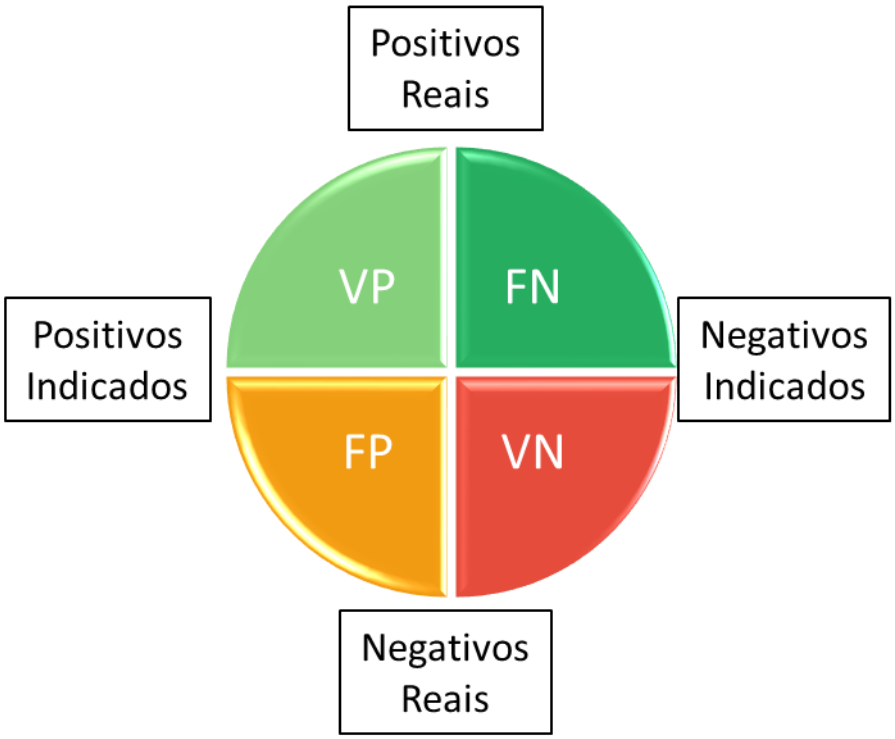


Figura 4 - Matriz de confusão

Esta representação dos resultados é preferível relativamente a uma simples Taxa de Acerto, dado essa poder facilmente induzir em erro: caso os dados de treino não sejam balanceados, ou seja, tenham uma densidade desproporcionada de classes, um classificador facilmente pode ficar enviesado para classificar todas as amostras como uma só classe.

Considere-se, por exemplo, o caso de uma amostra de 95 maças e 5 laranjas. Um classificador, face a esta amostra, pode ficar enviesado por forma a classificar todos os exemplos como maças. Dessa forma, teria uma Taxa de Acerto de 95%, sendo, no entanto, incapaz de distinguir uma laranja de uma maça.

No entanto, recorrendo outras métricas baseadas na matriz de confusão, é possível ter uma avaliação mais correta do desempenho de um classificador:

- Precisão: Proporção de exemplos de uma classe corretamente classificados face a todos os elementos classificados como sendo dessa classe. Uma precisão de 80%, por exemplo, significa que 80% dos elementos classificados como maçãs serão efetivamente maçãs, sendo os outros 20% mal classificados.

$$Precisão = \frac{VP}{VP + FP}$$

- Sensibilidade: Proporção de exemplos de uma classe corretamente classificados face a todos os elementos reais da classe na amostra de teste. Uma sensibilidade de 60%, por exemplo, implica que, de todas as maçãs na amostra, 60% são classificadas como tal.

$$Sensibilidade = \frac{VP}{VP + FN}$$

- *F_1 score*: Média harmónica entre Precisão e Sensibilidade. Combina os dois indicadores anteriores num só.

$$F_1 score: = \frac{2 \times VP}{2 \times VP + FP + FN}$$

A Taxa de Acerto, anteriormente referida, representa a proporção total de classificações corretas, calculando-se pela seguinte fórmula:

$$Taxa\ de\ Acerto = \frac{VP + VN}{Amostra\ Total}$$

2.4.4 Validação cruzada

Validação cruzada (Gama et al. 2012) é uma técnica de reamostragem para a validação de um modelo. No processo de treino de um algoritmo preditivo, este irá ajustar os seus parâmetros por forma a melhor explicar os dados com que foi alimentado. Pode suceder que o modelo fique demasiado ajustado aos dados de treino, não conseguido prever corretamente a classe de um novo elemento. Este sobreajustamento denomina-se de *overfitting* e é particularmente provável de ocorrer quando o conjunto de dados de treino é reduzido ou quando o número de atributos do modelo é elevado (Kohavi and Sommerfield 1995). O objetivo da Validação Cruzada consiste, assim, em remover o viés associado ao *overfitting* do classificador, bem como contrabalançar a aleatoriedade de amostras pequenas, por forma a conseguir avaliar-se corretamente o desempenho de um classificador. Tal é conseguido testando múltiplas amostragens para os dados de treino e calculando a média dos resultados, havendo vários métodos de validação cruzada.

K-fold cross-validation

No método de validação cruzada *K-fold*, o conjunto original irá ser dividido em *K* subgrupos de tamanho idêntico. Cada um dos *K* conjuntos será, à vez, usado como dados de teste, sendo os restantes *K-1* grupos usados como dados de treino. O modelo será então treinado *K* vezes, cada vez com uma amostra diferente como dados de teste até todos os subgrupos terem sido usados, sendo os respetivos resultados registados em cada iteração. Calcula-se, por fim, a média dos resultados obtidos, utilizando esses valores como performance do classificador.

A vantagem deste método sobre a repetição de testes com amostragem aleatória assenta no facto de todas as observações serem tanto para teste como para validação, sendo cada observação testada exatamente uma vez.

Stratified K-fold cross-validation

Na validação cruzada estratificada, cada um dos K grupos de dados mantém a mesma proporção de elementos de cada classe que a população original, por forma a não criar amostras bastante desbalanceadas.

2.5 Aplicação de KDD à área da Saúde

Da pesquisa realizada, a aplicação de técnicas de KDD e *Data Mining* a dados clínicos já tem vindo a ser realizada desde há alguns anos. Um dos seus usos foi já, por exemplo, a estimação do tempo de cirurgias em bloco operatório, por forma a otimizar o agendamento de cirurgias (C. G. Gomes 2014) e (Sperandio 2015), usando por base, na análise, dados de instituições portuguesas.

A previsão de reinternamentos clínicos com recurso a técnicas de *Data Mining* tem também vindo a ser explorada, havendo já tentativas de criar modelos que prevejam qualquer tipo de reinternamento, usando grandes quantidades de dados recolhidos de hospitais americanos

(Zheng et al. 2015), bem como o cálculo do tempo esperado para a readmissão de doentes (Helm et al. 2013).

No entanto, a maioria dos trabalhos encontrados dentro da área são relativos à previsão de reinternamentos de patologias específicas, como (Zolfaghar et al. 2013). Restringindo o universo em análise a uma só condição, possibilita a incorporação de *inputs* especificamente relevantes para a condição em análise, bem como dados de análises clínicas relevantes, o que, por norma, conduz a resultados superiores, apesar de reduzir o espectro de aplicabilidade.

3 Metodologia aplicada na Previsão da ocorrência de Reinternamentos

Ao longo do desenvolvimento do projeto da tese, a metodologia seguida foi, como já referido, a metodologia CRISP-DM. Ao longo deste capítulo será detalhado os passos seguidos em cada parte do processo. Para além disso, será fundamentada cada decisão tomada.

3.1 Compreensão da área de negócio

O projeto desenvolvido envolve, por forma a conseguir modelar corretamente a previsão de reinternamentos, a compreensão da problemática dos reinternamentos de um ponto de vista da saúde. Tal envolve, numa fase inicial, a análise de dados empíricos e estudos já existentes sobre previsão de reinternamentos, bem como tentativas prévias de modelação do problema.

Outro fator importante será identificar e representar corretamente a informação relativa a cada episódio de internamento da forma mais relevante para o algoritmo. Para tal, foi recorrida à opinião de profissionais da área da saúde (farmácia, medicina e enfermagem) por forma a melhor integrar os dados relativos a diagnósticos clínicos, administrações de medicamentos, bloco operatório e serviço de alta.

De acordo com a Organização Mundial da Saúde (OMS), doenças crónicas, tais como doenças cardiovasculares, a diabetes, a obesidade, o cancro e as doenças respiratórias correspondem a cerca de 59 por cento do total de 57 milhões de mortes por ano e 46 por cento do total de doenças (OMS 2005).

Num levantamento preliminar do perfil associado a casos de reinternamento, constatou-se ser, em termos de patologias, condições crónicas, como cancros, *Human Immunodeficiency Virus* (HIV), insuficiência cardíaca ou Doença Pulmonar Obstrutiva Crónica (DPCO), propensas a recaídas. O risco de reinternamento, segundo as mesmas fontes (OMS 2005), é mais elevado em pessoas idosas, pessoas com comorbilidades, como diabetes, tabagismo, colesterol elevado, etc., pessoas com fracos cuidados de saúde em casa, que vivem sozinhas, com baixos rendimentos ou sem acesso a medicação.

Um episódio de internamento define-se como a estadia de um utente num hospital de agudos, iniciando-se com a sua admissão, e terminando com a alta hospitalar (Lopes 2009).

Com o objetivo de mapear, com clareza a informação de cada episódio, foi elaborado um esquema relacional da informação considerada no modelo (Figura 5), por forma a realizar extração de dados da base de dados hospitalar:

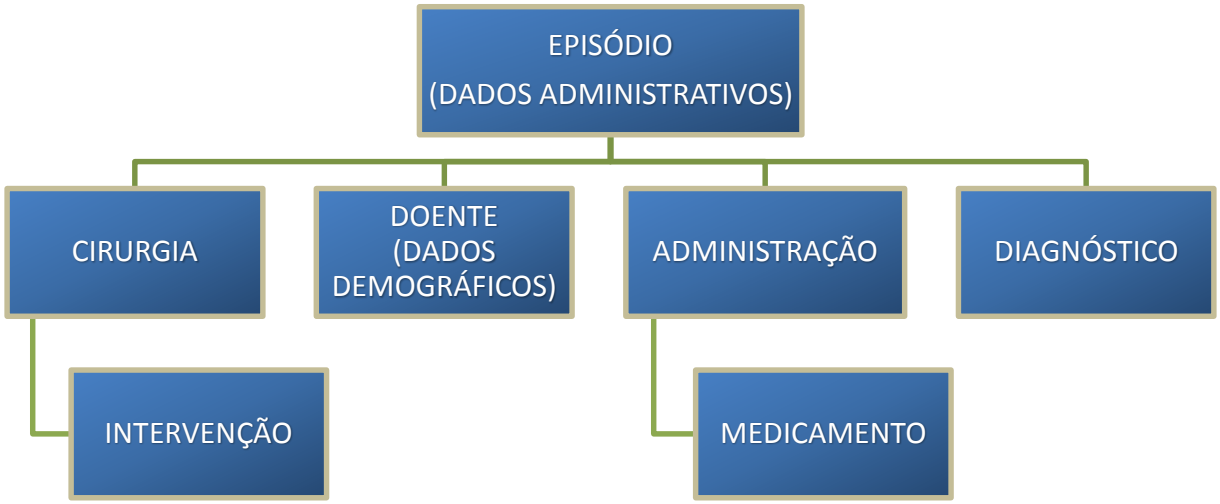


Figura 5 - Esquema relacional da informação

3.2 Recolha de dados

A recolha de dados foi feita através de *queries* em SQL, sendo a fonte uma base de dados multidimensional em *SQL Server*. Os dados obtidos foram exportados para ficheiros CSV e posteriormente incorporados num projeto de *Knime Analytics Platform*® (Knime), ferramenta utilizada no projeto.

Os campos extraídos foram escolhidos por forma a incorporar todos os elementos que, de acordo com a informação empírica, poderiam ser relevantes na previsão de casos de reinternamento. Tal resultou num conjunto de tabelas ligadas através do *id* de episódio, representadas na Figura 5.

Apresenta-se, de seguida, o conteúdo das tabelas, em termos gerais, sendo referidos os campos com principal interesse de cada uma. Não serão enumerados todos os campos presentes pela dimensão elevada das tabelas, sendo que muitos deles foram posteriormente descartados.

As tabelas referidas englobam os seguintes dados:

- Episódio - dados administrativos do episódio: dia e hora de admissão e alta; serviço e valência de admissão e alta; tipo de episódio-pai (Urgência, Consulta, etc.), prioridade segundo a triagem de Manchester (para doentes vindos da Urgência), GDH e respetiva Grande Categoria de Diagnóstico (GCD); Note-se que estes dois últimos campos são atribuídos algum tempo após o final do episódio, não podendo, por essa razão, ser incluídos numa análise preditiva aquando da alta clínica;
- Doente – dados demográficos relativos aos doentes: idade, residência, naturalidade, nacionalidade estado civil, escolaridade, entre outros campos;
- Cirurgia – informação do bloco operatório relativa a cirurgias realizadas: dia e hora de início e fim da cirurgia, anestesia usada e duração da mesma, médico que a realizou, tipo de cirurgia, urgente ou eletiva;
    - Intervenção – cada cirurgia envolve uma ou mais intervenções: tipo de intervenção, codificado em *International Classification of Diseases*, 9ª revisão (ICD-9), data e hora de início e fim;

- Administração – dados da administração de medicamentos a cada doente: dia e hora de administração, qual o medicamento administrado e respetiva quantidade, bem como as unidades em que foi administrado;
    - Medicamento – dados dos medicamentos: nome, princípio ativo, dosagem, unidades; Note-se que cada administração corresponde a um só medicamento;
- Diagnóstico – dados de diagnósticos atribuídos a cada paciente: podem ser o diagnóstico principal ou um dos secundários, codificados em ICD-9.

Esta recolha foi propositadamente abrangente (cerca de 120 campos), por forma a incluir todos os campos que pudessem ajudar na previsão. O conteúdo destes foi posteriormente analisado em *Microsoft Excel*, através de *pivot tables* e *power pivot tables* para as tabelas de maiores dimensões, por forma a detetar problemas de qualidade e analisar o domínio de cada campo. Muitos dos campos foram descartados por estarem muito incompletos ou ter dados não relevantes no seu conteúdo (domínio com apenas um valor, *flags* administrativas ou campos de texto livre).

3.3 Tratamento de dados

A principal ferramenta utilizada no desenvolvimento deste projeto foi o *Knime Analytics Platform*®, pelo seu interface gráfico intuitivo e grande capacidade de processamento de dados. O Knime possui uma extensa biblioteca de funções que facilita tanto o processamento e manipulação dos dados, bem como a implementação dos algoritmos preditivos e realização de testes customizáveis. Para além disso, permite a incorporação de funções em Weka e código em R (ver Anexo A), ambos utilizados no projeto.

3.3.1 Construção do *dataset*

Após a recolha e importação dos dados para o projecto em Knime, foram realizadas diversas operações de limpeza de dados, tais como correções de caracteres em campos de texto devido a erros humanos de *input*, problemas de *encoding* com caracteres portugueses, espaços em branco, sinais de pontuação nos campos, entre outros. Outras operações envolvem definição do tipo de variável de cada campo (*string, integer, single, boolean)* e conversão de campos de texto em datas, para as datas de admissão e alta (ver Anexo B).

Após estas operações, foi realizada a junção de tabelas pelo *id* de episódio, obtendo, dessa forma, um *dataset* que incluísse toda a informação relevante para realizar a previsão numa só linha da tabela, correspondendo, cada linha, a um episódio distinto.

Neste processo de junção muitos episódios não tiveram correspondência noutras tabelas. A ausência de dados pode ocorrer devido a duas razões: inexistência real da informação ou registos incompletos de dados. No primeiro caso, os dados estão corretos e representam a realidade do episódio. Pode ocorrer, por exemplo, um episódio sem realização de cirurgias ou toma de medicamentos. No entanto, episódios sem um doente ou diagnóstico associado são registos incompletos, tornando o episódio inutilizável por falta de dados.

Os registos podem encontrar-se incompletos devido a erros ou lapsos de preenchimento, por se tratarem de dados importados de outras soluções de armazenamento de dados previamente utilizadas no hospital, ou por, em alguns Serviços do Hospital, não ser utilizada a solução da Glintt HS.

Finalmente, refira-se ainda um passo adicional de processamento relativamente ao *dataset* para alimentar o MLP: pelo seu funcionamento, os *inputs* deste algoritmo devem, idealmente, ser binários. Para tal, realizaram-se os seguintes passos:

- Campos qualitativos, tal como o Serviço de Alta, foram discretizados em colunas binárias: foi acrescentado ao *dataset* uma coluna para cada valor possível do campo, possuindo um valor de 1 ou 0 consoante o valor dessa coluna no episódio em questão;
- Campos numéricos, tal como a idade ou o número de dias de internamento foram discretizados em grupos de valores. Os intervalos para esses grupos foram calculados a partir da distribuição dos valores do campo, por forma a obter grupos o mais homogéneos possível.

Esses grupos sofreram, posteriormente, a mesma discretização binária que os campos qualitativos, por forma a obter, finalmente, um *dataset* integralmente binário.

3.3.2 Filtragem de Dados

Por forma a minimizar o impacto de dados errados ou incompletos, foram estabelecidos um conjunto de requisitos mínimos dos dados para serem aceites pelo modelo, com o objetivo de melhorar a fiabilidade dos resultados.

Um dos critérios estabelecidos foi o episódio ter, pelo menos, um diagnóstico associado, por forma a caracterizar a patologia do doente internado. Sem este campo, de um ponto de vista lógico, é impossível estabelecer um critério de comparação relativamente a outros episódios.

O segundo critério considerado foi o episódio ter a data de admissão superior a 1/1/2012 e a data de alta inferior a 31/12/2014. Dados antigos (anteriores a 2012), apresentavam-se bastante incompletos, principalmente devido a ausência de diagnósticos associados a cada episódio. A partir do ano 2012 (inclusive), verificou-se que os diagnósticos se encontravam, maioritariamente, devidamente preenchidos. Dessa forma, apenas foram considerados para análise de dados a partir deste ano.

Para além disso, dados bastante recentes (2015) também não foram considerados para esta análise, uma vez que (1) muitos não têm ainda um diagnóstico principal associado, atendendo a que este apenas é atribuído algum tempo após alta e (2), apenas após um período de 30 dias da alta se pode saber se o episódio efetivamente resultou, ou não, num reinternamento.

Por fim, optou-se por descartar episódios que tivessem resultado em falecimento, uma vez que estes casos induziriam em erro o algoritmo. Tal ocorreria por (1) nunca poderiam resultar em reinternamentos, sendo categorizados, para o algoritmos, como não-reinternamentos e (2) o perfil de um episódio que resulta em falecimento é fundamentalmente diferente de um que resulte num não reinternamento, pelo que seria falacioso manter esses episódios no *dataset*.

3.3.3 Cálculo de métricas relevantes para o modelo

Cálculo do campo "Gerou Reinternamento"

Por forma a identificar os episódios que geraram reinternamento, incluiu-se no *dataset* uma "coluna objetivo" que identificasse a classe de cada episódio: "gerou reinternamento" ou " não gerou reinternamento".

Esta coluna deve ter indicado se esse episódio gerou ou não um episódio subsequente de reinternamento. Os critérios considerados para tais são os mesmos da faturação hospitalar, ou seja, ser um novo episódio de internamento, do mesmo paciente (mesmo *id*) cujo GDH atribuído esteja dentro da mesma GCD.

Com o objetivo de obter este campo, recorreu-se à elaboração de uma *query* em SQL que tentasse encontrar, para cada episódio de internamento, se existe um outro, nos 30 dias após alta, com o mesmo doente e o mesmo GCD.

Cálculo de Medicamentos tomados em cada episódio

Face à panóplia de dados disponíveis relativamente à administração medicamentosa, tentou-se perceber qual seria o indicador mais adequado e que contribuísse para a obtenção de resultados fiáveis e representativos do modelo, ajudando a explicar um caso de reinternamento.

Neste sentido, optou-se por, primeiramente reduzir a dimensão do campo considerando apenas os princípios ativos de cada medicamento, reduzindo, assim de cerca de 850 medicamentos distintos para cerca de 450 princípios ativos. Posteriormente, procedeu-se ao teste das seguintes hipóteses relativamente à forma como os dados iriam ser incorporados no modelo:

- Simples contagem do número de princípios ativos tomados no episódio;
- Indicação binária da toma de um princípio ativo no episódio (introduzindo uma coluna por princípio ativo);
- Indicação do número de dias que tomou cada princípio ativo durante o reinternamento (uma coluna por princípio);
- Indicação da quantidade total administrada por princípio ativo.

Os primeiros três indicadores supramencionados foram testados no modelo. No entanto, devido a unidades não SI e pouco estandardizadas relativamente às quantidades administradas de cada medicamento, não foi possível calcular o último indicador.

Cálculo de Diagnósticos totais atribuídos a cada episódio

Dado que, por episódio poderia haver até cerca de 30 diagnósticos distintos foi necessário escrutinar diferentes indicadores para que, de forma análoga aos medicamentos, se tentasse perceber qual seria o mais adequado para melhorar os resultados do classificador.

Os indicadores testados foram os seguintes:

- Simples contagem do número de diagnósticos secundários associados ao episódio;
- Indicação binária de ter ou não um diagnóstico específico associado (introduzindo uma coluna por diagnóstico); Note-se, no entanto, que, ao todo, existem cerca de 1150 diagnósticos únicos presentes na amostra, o que aumenta consideravelmente a dimensão e, consequentemente, o tempo necessário para processamento da tabela.

Ambos os indicadores suprarreferidos foram testados no modelo. Adicionalmente, com vista a reduzir a dimensão do universo de diagnósticos, de forma a manter a representatividade funcional da atribuição de um diagnóstico, recorreu-se a uma ferramenta de classificação clínica de diagnósticos ICD-9.

A ferramenta em questão *Clinical Classifications Software* (CCS), desenvolvida pela *Healthcare Cost and Utilization Project* (HCUP), (HCUP 2015) tem como objetivo agrupar os milhares de diagnósticos e procedimentos do ICD-9 em patologias e categorias de procedimentos, reduzindo significativamente a dimensão destes campos de forma funcional.

A ferramenta estabelece uma correspondência de vários diagnósticos para uma só patologia, sendo que cada diagnóstico corresponde a uma só patologia. Apresenta-se, em seguida, dois exemplos dessas correspondências (Tabelas 1 e 2):

Tabela 1 - Correspondência CCS (exemplo 1)

Diagnósticos	Patologia Correspondente
1550 1551 1552 2308 V1007	16 *Cancer of liver and intrahepatic bile duct*

Tabela 2 - Correspondência CCS (exemplo 2)

Diagnósticos	Patologia Correspondente
1570 1571 1572 1573 1574 1578 1579	17 *Cancer of pancreas*

Desta forma, testou-se ainda mais um indicador, baseado nas correspondências da ferramenta mencionada:

- Indicação binária de ter ou não uma patologia específica associada (introduzindo uma coluna por princípio ativo); esta transformação reduziu a dimensão do campo para cerca de 300 valores possíveis.

3.3.4 Estudo Estatístico de Atributos

Por forma a discernir se os atributos considerados no *dataset* contribuem efetivamente para a previsão de casos de reinternamento, bem como saber quais são os atributos que mais impacto têm sobre a taxa de reinternamento é necessário analisar a distribuição dos dados relativamente às 2 classes em análise.

Para tal recorreu-se, uma vez mais à funcionalidade de *pivot tables* do Excel, utilizando-a para visualizar, para cada atributo, a percentagem de reinternamentos relativamente a cada valor possível do seu domínio.

O critério de análise considerado para a inclusão ou descarte de campos, bem como para avaliar o impacto do atributo foi a distribuição dos dados relativamente à taxa média de reinternamentos.

Assim, considerando os episódios divididos de acordo com os valores possíveis para um atributo, foi comparada a taxa de reinternamentos de cada grupo com a taxa média. Caso o atributo seja irrelevante para distinguir um episódio que gere um reinternamento de um que não gere, em princípio, a taxa de reinternamento para cada valor será igual, ou muito próxima da média. No entanto, caso o atributo seja relevante na distinção, para cada valor do domínio, a taxa de reinternamentos deverá alterar significativamente, refletindo o impacto da variável sobre esta.

Esta análise estatística realizada sobre as variáveis trata-se, em si, de informação relevante para discussão, sendo, por essa razão, os dados obtidos apresentados no capítulo seguinte em maior detalhe.

Os gráficos apresentados (Figuras 6 a 12), serão, no entanto, apenas relativamente aos atributos mais relevantes, pela dimensão que tomaria incluir todos os atributos. Dessa forma, apresentar-se-ão os gráficos da taxa de reinternamento relativamente a:

- Idade;
- Género;
- Patologia;
- Faixa etária;

- Especialidade;
- Mês;
- Número de episódios anteriores do mesmo doente (campo calculado).

3.4 Modelação

3.4.1 Modelo Inicial

Uma vez completo o *dataset* final, este foi utilizado para alimentar vários modelos preditivos, por forma a determinar aquele que melhores previsões consegue obter.

Para tal, foram testados os seguintes algoritmos:

- *Random Forest*
- MLP
- SVM
- *Naive Bayes*
- *Adaboost* com árvores de decisão

Os algoritmos testados mais exaustivamente foram o RF e o MLP, uma vez que, sendo o *dataset* bastante semelhante ao testado por (Pereira 2014), era já esperado, ter como ponto de partida para melhoria os resultados do RF. Alguns testes preliminares corroboraram essa expectativa, tendo os restantes algoritmos obtido resultados notoriamente inferiores.

Para cada um destes algoritmos foi testado, para além de um conjunto de *inputs* transversal a todos, os dados relativos às administrações e diagnósticos, em vários indicadores, de acordo com as hipóteses previamente apresentadas. Devido ao elevado tempo de processamento associado ao teste de *datasets* com um grande número de dimensões, estes testes foram realizados separadamente, por forma a discernir se a adição dos dados, em cada forma, traz melhorias aos resultados das previsões.

3.4.2 Modelo Específico - Análise de Pneumonias

A abordagem inicial ao problema consistiu, como já referido, em tentar obter um modelo que conseguisse prever qualquer caso de reinternamento, independentemente da patologia em questão.

Tal implica, por questões de completude da informação, a inclusão de bastantes campos que apenas apresentam valores relevantes num número reduzido de casos. Dependendo da condição do doente, estes serão irrelevantes para outras previsões, o que pode dificultar o discernimento de padrões do algoritmo. Como tal, formulou-se a hipótese de considerar os dados relativos apenas a uma patologia específica, por forma a tentar filtrar apenas os campos relevantes para a mesma e, dessa maneira, melhorar os resultados das previsões.

Esta hipótese surgiu com base nos artigos (K. Kanel, Elster, and Vrbin 2010) e (K. T. Kanel, Elster, and Vrbin 2011), uma vez que estes tratam o reinternamento com base apenas numa patologia.

A abordagem de ambos os artigos é bastante semelhante. Desta forma, por motivos de síntese, descreve-se, de seguida, a metodologia seguida no artigo (K. Kanel, Elster, and Vrbin 2010).

O artigo consiste numa análise efetuada aos diagnósticos resultantes do internamento de doentes infetados com HIV, a causa do Síndrome de Imunodeficiência Adquirida (SIDA). O principal objetivo do artigo é fornecer informação sobre os doentes portadores do vírus da

SIDA (seropositivos) e os respetivos padrões relativamente à admissão e readmissão hospitalar. A análise foca-se essencialmente nas características dos doentes portadores de HIV, nos padrões de admissão e readmissão hospitalar e, por fim, na oportunidade de reduzir a taxa de readmissão destes doentes.

Os autores analisaram os diagnósticos principais e secundários (ou comorbilidades) mais frequentes que levam ao internamento de doentes que já tenham histórico de diagnóstico principal ou secundário de infeção por HIV. Para além disso, também tiveram o cuidado de averiguar a prevalência de diagnósticos relacionados com outras condições de saúde (ou seja, diagnósticos não associados diretamente à condição patológica em análise), tais como: depressões, dependência de substâncias, etc.

Aos doentes diagnosticado o vírus da SIDA, listaram os dez diagnósticos (independentemente de serem principais ou secundários) que originam mais frequentemente episódios de internamento. Da tabela obtida, conseguiram averiguar, por exemplo, que o "Abuso de Substâncias Não Viciantes" liderava e representa mais de um quarto dos internamentos, o que tal é justificado pela má gestão medicamentosa do doente associada à dependência de tabaco e álcool.

Em segundo lugar encontrava-se o diagnóstico "Hepatite Viral", o qual estava presente em quase 25% dos doentes admitidos. Tal é explicado pelo facto de cerca de 30% dos doentes infetados com o vírus da SIDA, nos EUA, estarem co- infetados com outros vírus como a Hepatite C e Hepatite B. Para além disso, um facto surpreendente para os autores, foi a presença do diagnóstico "Depressão" no top 10 dos diagnósticos mais comuns. Neste sentido, cruzaram os doentes com a comorbilidade "Depressão", comorbilidade "Abuso de Substâncias Não Viciantes" e ambas, com os seis diagnósticos mais comum da tabela anterior. Como exemplo, verificaram que mais de 50% dos doentes com o diagnóstico "Hepatite Viral" têm ambos os diagnósticos "Depressão" e "Abuso de Substâncias Não Viciantes" (K. Kanel, Elster, and Vrbin 2010).

Caso de estudo - Pneumonias

Atendendo ao artigo (Campos 2014), Portugal é o país da União Europeia com a maior taxa de mortalidades por pneumonia em hospitais. Neste sentido, achou-se pertinente focar esta componente do projeto na análise dos episódios diagnosticados com pneumonia.

De forma análoga à abordagem adotada no artigo (K. Kanel, Elster, and Vrbin 2010), tentou-se determinar os diagnósticos secundários que mais afetam a taxa de reinternamento de doentes com pneumonia.

Para tal, foi calculado, para cada diagnóstico secundário associado a casos diagnosticados com pneumonia, a taxa de reinternamentos da população com o diagnóstico secundário e da população sem o mesmo diagnóstico.

De seguida, efetuou-se a subtração entre as mesmas duas taxas, por forma a determinar o conjunto dos diagnósticos cuja presença num dado episódio mais afeta a probabilidade de este resultar num reinternamento.

Também houve o cuidado de realizar a mesma análise relativamente às patologias obtidas através da ferramenta CCS da HCUP, de forma análoga.

Finalmente, foi definido um valor mínimo para essa diferença (15%), bem como um número mínimo de ocorrências de episódios com ambos os diagnósticos, pneumonia e o secundário (10 casos) para determinar os 20 diagnósticos mais relevantes.

Por forma a enquadrar os diagnósticos secundários encontrados com os fatores de risco relacionados com a pneumonia foi necessário recorrer a fontes de informação adequadas.

De acordo com (L. Gomes 2001), os fatores de risco que exponenciam o aparecimento da patologia de pneumonia são:

- Idade - pessoas com mais de 65 anos, dado que sofrem de alterações no sistema imunológico e estão mais expostas a determinadas doenças do foro psiquiátrico, tal como demência, alzheimer, etc.;
- Tabagismo – o tabaco potencia o desenvolvimento da DPOC;
- Insuficiências cardíacas – aumentam cerca de duas vezes a probabilidade de contrair pneumonia;
- Imunossupressão – existem comorbilidades mais relacionadas com a imunodisfunção local ou sistémica, tais como DPOC ou *diabetes mellitus* ou com a imunodepleção grave, tais como infeção por HIV, SIDA, doentes transplantados ou neutropénicos;
- Gripes – este fator está intimamente ligado com a idade, dado que são os idosos os mais afetados pelo vírus da gripe. Adicionalmente, destaca-se as condições atmosféricas e a virulência do vírus como fatores determinantes para o aumento da incidência da gripe.

Estes fatores de risco encontram-se intimamente relacionados com o surgimento de mais ou menos comorbilidades que, por sua vez, são representados através de diagnósticos secundários ICD-9. Estes, no entanto, não abrangem todo o espectro de comorbilidades relevantes para a previsão de casos de reinternamento. Apesar disso, são o indicador mais fidedigno para representar a condição do doente ao longo da sua estadia hospitalar.

Para além dos diagnósticos secundários, foram ainda analisados os princípios ativos dos medicamentos administrados a este universo. Estes deverão refletir o tratamento tanto da pneumonia como das restantes comorbilidades, ou seja, dos diagnósticos secundários.

A metodologia para determinar os princípios ativos que mais afetam a taxa de reinternamento foi análoga à adotada para analisar os diagnósticos secundários, obtendo os 20 princípios mais relevantes.

3.5 Avaliação de Modelos

Na avaliação da qualidade de cada solução obtida, foram analisados os resultados relativamente a três métricas distintas: Precisão, Sensibilidade e Taxa de Acerto.

Atendendo a que, no caso em análise, a prioridade é conseguir prever casos de reinternamento, é necessário realçar que é mais importante prever com fiabilidade a classe positiva do que a negativa, pelo que os indicadores mais relevantes de maximizar serão: em primeiro lugar, a sensibilidade, por forma a captar o máximo possível de casos de reinternamentos, aumentando a abrangência do algoritmo; em segundo lugar, a precisão, por forma a obter previsões o mais rigorosas possível.

4 Apresentação de Resultados

Nesta secção pretende-se apresentar e interpretar os resultados obtidos nos modelos testados.

A população analisada no modelo geral consta de 36396 episódios distintos, que cumprem os critérios de filtragem estabelecidos, tendo sido abordados em mais detalhe na secção 3.3.2:

- Episódios com início após 01/01/2012 (inclusive) e alta antes de 31/12/2014 (inclusive);
- Que tenham pelo menos um diagnóstico associado;
- Que não tenham resultado num falecimento.

Os episódios considerados têm uma taxa geral de 9.25% de reinternamento, com a seguinte distribuição por classes:

- 3365 episódios geraram reinternamento;
- 33031 episódios não geraram reinternamento.

4.1 Análise estatística dos atributos

Apresenta-se, de seguida, os gráficos relativos à análise da taxa de reinternamentos relativamente a cada atributo.

- **Género**

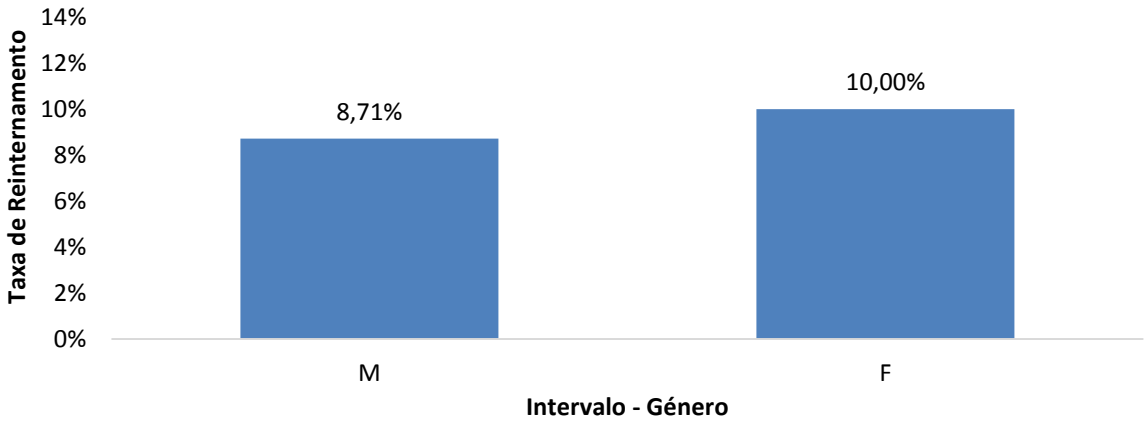


Figura 6 - Taxa de reinternamento Vs Género

O género (Figura 6) parece não influenciar muito a taxa de reinternamento, apesar de haver uma tendência para elementos do sexo feminino serem, de modo geral, reinternados um pouco mais frequentemente (1.29 pontos percentuais) do que do sexo masculino.

- **Patologia**

Top 10 de Patologias com maior Taxa de Reinternamento

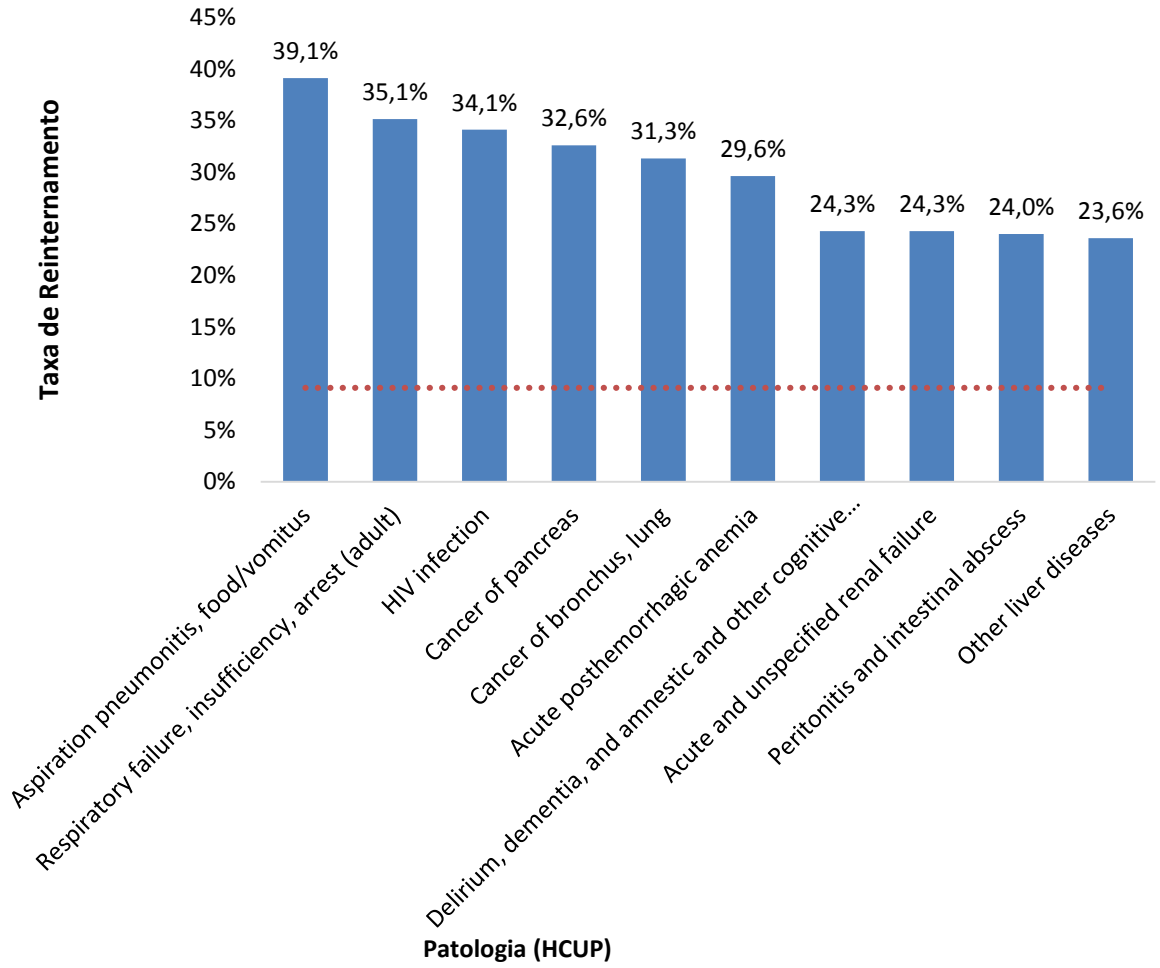


Figura 7 - Taxa de reinternamento Vs Patologia

Os dados apresentados na Figura 7 correspondem às 10 patologias com maior taxa de reinternamento, de acordo com a patologia atribuída pela ferramenta de agrupamento de diagnósticos ICD-9 da HCUP (HCUP 2015).

Estes dados apenas incluem patologias com mais de 20 ocorrências nos 3 anos considerados, por forma a terem um número significativo de ocorrências Não são apresentadas todas as patologias devido à dimensão das mesmas (ao todo, existem 249 patologias distintas na população). É, no entanto, evidente, pelo afastamento relativamente à média (9.25%), representada pela linha vermelha no gráfico que a patologia a que o diagnóstico atribuído corresponde se trata de um fator determinante na diferenciação de episódios que resultam em reinternamentos.

A taxa de reinternamento por patologia vai desde 0.00%, para patologias como entorses, reações alérgicas ou feridas expostas em extremidades dos membros até 39.1% para pneumonia por aspiração.

Note-se que, de entre as patologias apresentadas, bem como os elementos que se seguiriam, destacam-se patologias de:

- Especialidade médica de Gastrenterologia e especialidade cirúrgica de Cirurgia Geral (principalmente cancros);
- Especialidades médicas de Cardiologia e Pneumologia e cirúrgica Cardiotorácica (doenças cardiorrespiratórias, como paragem cardíaca, e cancros);
- Especialidade médica de Nefrologia e Cirúrgica de Urologia (doenças renais agudas e crónicas e infeções do trato urinário);
- Especialidade médica de Psiquiatria (demências, delírios e alterações de humor).

Top 10 de Grupos de Patologias com maior Taxa de Reinternamento

Para além da classificação de diagnósticos por patologia, a HCUP disponibiliza uma ferramenta de classificação adicional, multinível (quatro níveis), mais detalhada, disponível na mesma página web. Utilizando o primeiro nível dessa classificação, mais geral, com 17 valores possíveis, ao todo, foi construído o gráfico seguinte:

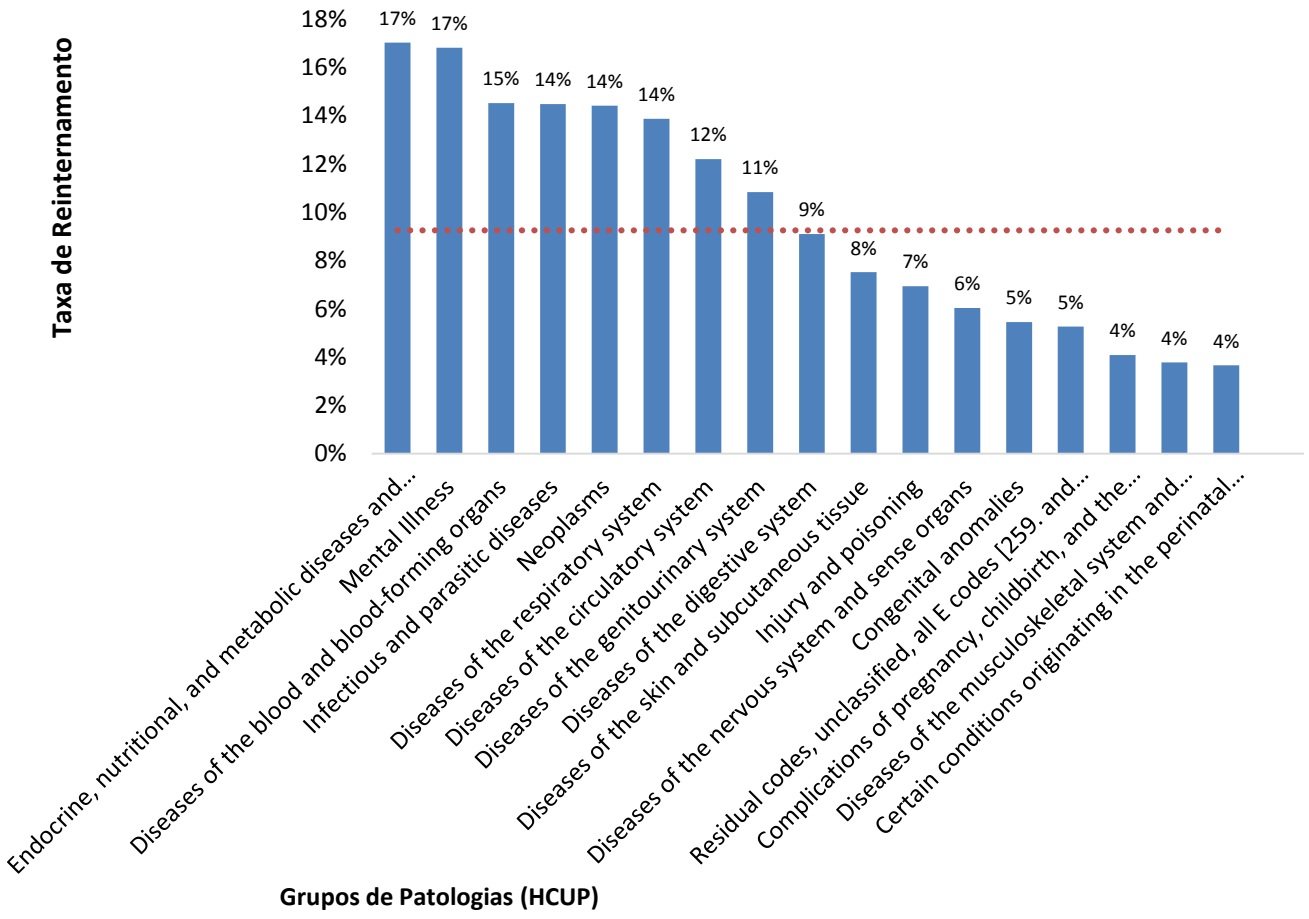


Figura 8 - Taxa de reinternamento Vs Grupos de Patologias

Contrariamente ao gráfico anterior (Figura 7), a Figura 8 apresenta os dados da população completa, ordenados por ordem decrescente de taxa de reinternamento. Apesar de não apresentar taxas tão afastadas da média quanto o gráfico anterior, que possui maior nível de

detalhe, é igualmente possível visualizar o efeito do diagnóstico sobre a probabilidade de reinternamento.

Dos grupos de patologias com maior taxa de reinternamento, destacam-se "Alterações imunológicas e patologias endócrinas e metabólicas" (17.03%) e "Patologias psiquiátricas" (16.82%).

Com menor taxa de reinternamento, destacam-se "Patologias do sistema músculo-esquelético" (3.78%) e "Condições geradas durante o período perinatal" (3.66%).

- **Faixa etária**

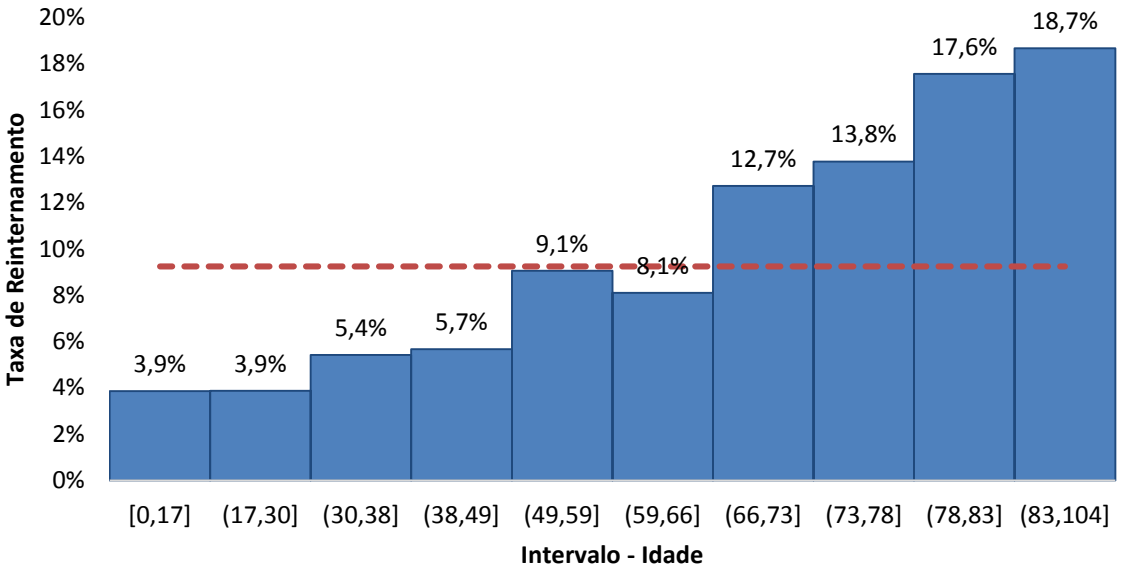


Figura 9 - Taxa de reinternamento Vs Idade

Como se pode constatar pela Figura 9, as idades dos doentes internados vão de 0 a 104 anos. Os intervalos apresentados contêm aproximadamente 3000 elementos cada, à exceção do primeiro grupo, que contém 6428 episódios.

É possível discernir um aumento do risco de reinternamento com o aumento da idade, o que está de acordo com as expectativas relativamente a este atributo.

Por, fim, é de realçar que a taxa de reinternamento até aos 50 anos se mantém relativamente estável.

- **Especialidade**

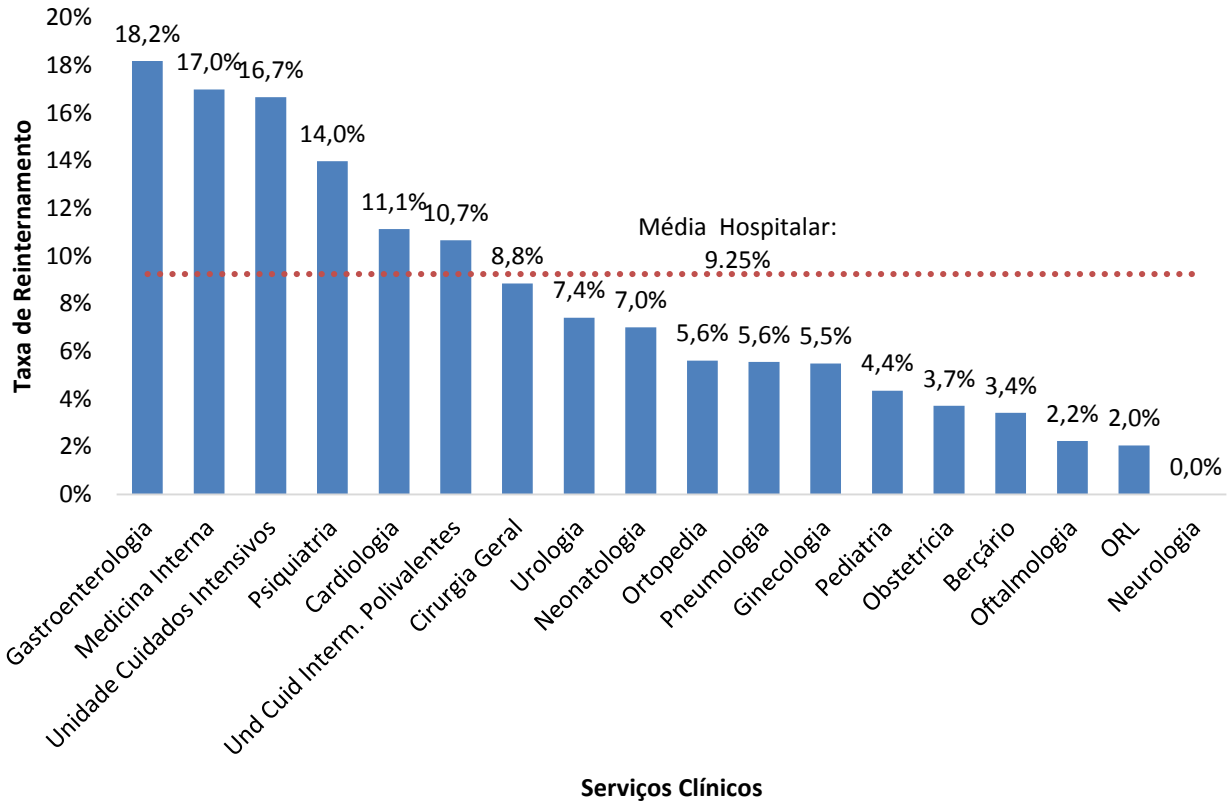


Figura 10 - Taxa de reinternamento Vs Especialidade

Analisando a taxa de reinternamento por especialidade (Figura 10), é possível concluir que este atributo é, também, bastante discriminante. Entre as especialidades com maior taxa de reinternamento, encontram-se Gastroenterologia, Psiquiatria e Cardiologia, tendo já sido indiciado pela análise de patologias que seriam elevadas.

No entanto, realça-se a taxa elevada das especialidades de Medicina Interna (17.0%) e dos Cuidados Intensivos (16.7%).

Na especialidade de berçário, tal como já indiciado pelo gráfico de Idade, a taxa de reinternamentos é baixa.

No gráfico, destaca-se ainda os 0% correspondentes à especialidade de Neurologia, podendo afirmar-se que todos os casos desta especialidade resultam ou numa recuperação ou num falecimento, de forma binária. Tal facto poderá ter a ver com casuística dos episódios analisados.

- **Mês**

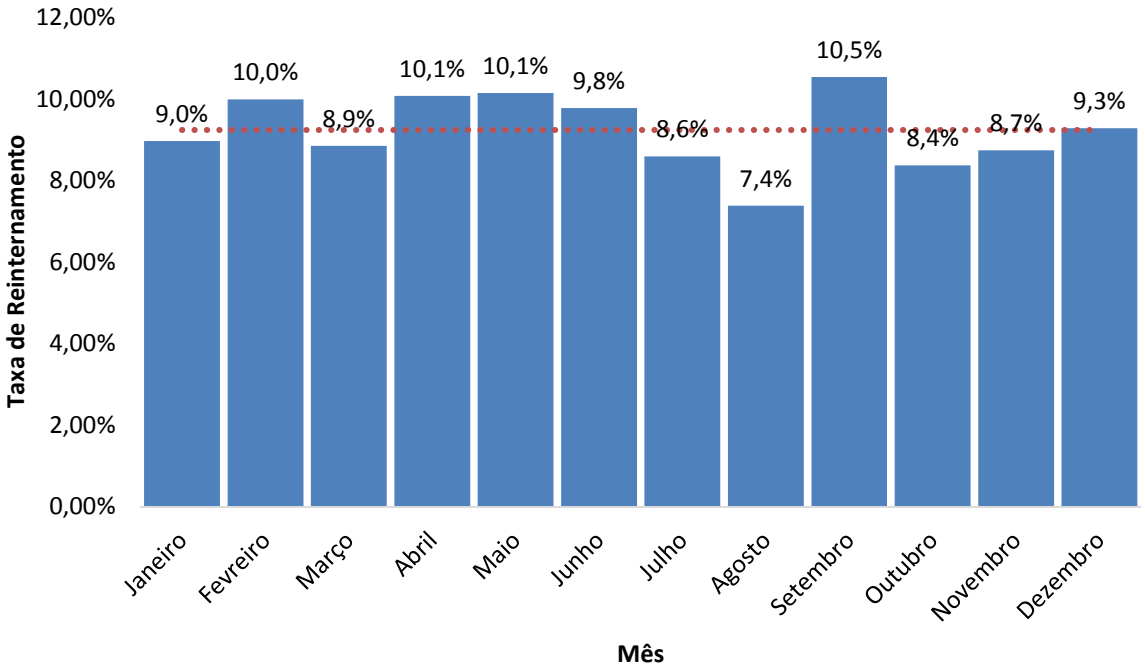


Figura 11 - Taxa de reinternamento Vs Mês

Como se pode verificar pela Figura 11, a taxa de reinternamentos mantém-se relativamente estável ao longo do ano. No entanto, apresenta um pico inferior em agosto (7.4%), seguido de um pico superior em setembro (10.5%). Tal poderá dever-se ao adiamento de cirurgias e internamentos consequentes devido ao período de férias entre julho e agosto.

- **Reinternamentos por número de episódios anteriores**

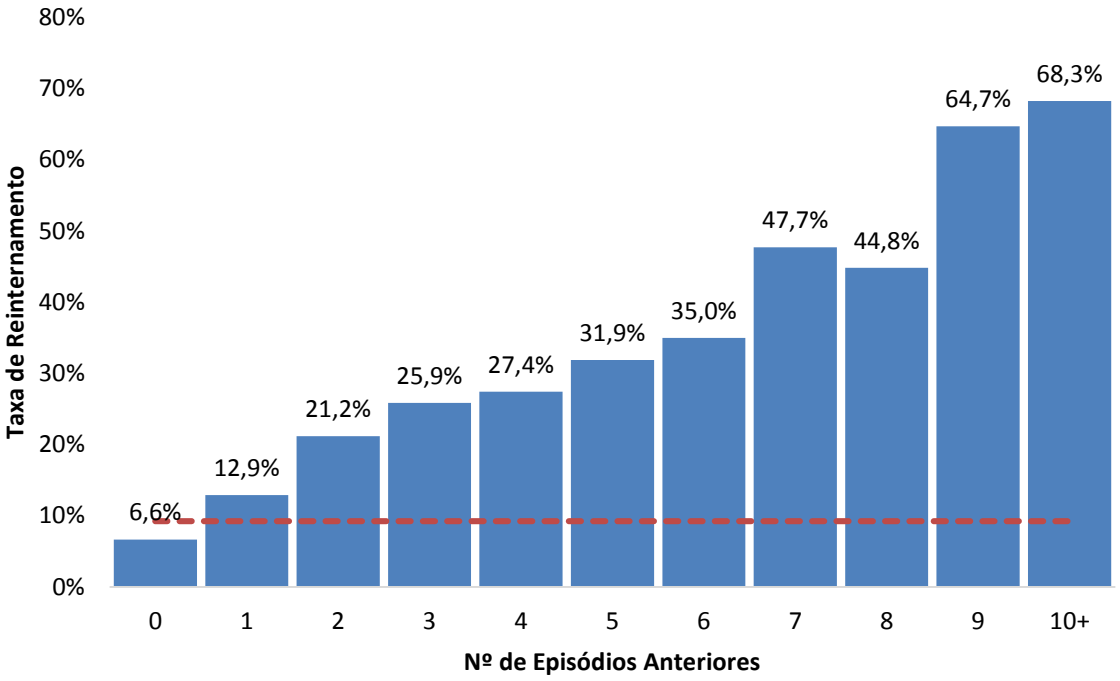


Figura 12 - Taxa de reinternamento Vs Nº de Episódios anteriores

Através da Figura 12 observa-se que a taxa de reinternamento aumenta de forma quase linear com o aumento do número de episódios de internamento do mesmo doente anterior ao episódio em análise.

O campo em análise apresentado foi calculado com base nos registos do hospital em análise. Dessa forma, o número de episódios anteriores é relativo a reinternamentos no mesmo hospital.

É de salientar, que 27511 episódios do total da população têm zero episódios anteriores. Ainda assim, os restantes apresentam, com cada episódio prévio adicional, um acréscimo significativo de risco de reinternamento, tratando-se, dessa forma, de um fator de risco extremamente relevante.

Lista de atributos final

A lista de Atributos utilizada para o modelo final, após a análise de cada atributo e testes foi a seguinte:

Demogáficos

- Número de dias de internamento;
- Mês de internamento;
- Dia da semana de Alta;
- Sexo;
- Idade do doente na admissão (em anos);
- Classificação - triagem de Manchester;

Administrativos

- Especialidade em que foi admitido;
- Serviço e Valência de Alta;
- Tipo de Episódio-Pai;
- Número de episódios de internamento anteriores;

Diagnósticos

- Categoria da patologia do diagnóstico principal (CCS da HCUP);
- Categoria do procedimento principal (CCS da HCUP);
- Classificação patologia (CCS HCUP) nível 1 e 2;
- Classificação procedimento (CCS HCUP) nível 1;
- Número de Diagnósticos Secundários atribuídos;

Bloco

- Número de Intervenções cirúrgicas realizadas durante o internamento;

Administrações

- Número de Princípios ativos administrados durante o internamento.

4.2 Resultados do Modelo Geral

Apresenta-se, nesta secção, os melhores resultados para cada um dos modelos testados:

- Modelo RF Geral:

Tabela 3 - Resultados do modelo RF Geral

Precisão	Sensibilidade	Taxa de Acerto
80.01%	8.20%	90.30%

Tendo testado com múltiplas combinações de *inputs*, aumentando o número de árvores (testadas entre 100 e 1000 árvores) e variando o número de atributos aleatórios por árvore, estes foram os melhores resultados obtidos (Tabela 3).

Notou-se, aquando do teste uma grande variação dos resultados consoante os *inputs* alimentados (acrescentando e retirando atributos do *dataset*).

Verificou-se ainda uma grande suscetibilidade do método à adição de dados pouco relevantes para a previsão, que pioram os resultados muito significativamente, comparativamente ao MLP que consegue atribuir pesos menores aos *inputs* de um certo atributo, mantendo-se mais consistente nos resultados apresentados.

Foram conseguidos resultados com maior sensibilidade (até 14.71%) mas pioravam drasticamente a precisão (máximo de 20.32%), não sendo, por isso, incluídos. No entanto, não foi possível melhorar a sensibilidade para além do valor referido, quer por variação dos atributos alimentados, quer por alteração dos parâmetros do algoritmo, o que limitaria consideravelmente a aplicabilidade do modelo.

Por essa razão, o MLP foi testado mais extensivamente, tendo sido utilizado para todos os testes cujos resultados se apresentam de seguida.

- MLP com Princípios ativos:

Tabela 4 - Resultados do modelo MLP com Princípios ativos

Precisão	Sensibilidade	Taxa de Acerto
23.838%	17.396%	85.751%

- MLP com Diagnósticos Secundários:

Tabela 5 - Resultados do modelo MLP com Diagnósticos Secundários

Precisão	Sensibilidade	Taxa de Acerto
18.655%	17.03%	85.458%

Ambos os modelos testados pioram consideravelmente a qualidade das previsões com a inclusão de todos os campos resultantes da discretização tanto dos diagnósticos secundários, como das categorias das patologias (de dimensão menor), como dos princípios ativos, como se pode constatar pelas Tabelas 4 e 5.

Acredita-se que tal será devido à grande dimensão das tabelas analisadas, sendo que cada campo das tabelas apenas é relevante num número restrito de episódios, sendo contraproducente para discernir padrões nos restantes episódios.

Por essa razão, os campos referidos foram removidos do modelo, sendo os melhores resultados conseguidos resultantes do modelo MLP geral que apenas inclui a informação relativa a esses atributos de forma simplificada, através de contagens de diagnósticos secundários e princípios ativos administrados.

- Modelo MLP Geral:

Como referido, este foi o modelo mais explorado e testado extensivamente. Foi realizado um conjunto de testes iterativos variando vários parâmetros do modelo por forma a obter a melhor solução possível, bem como compreender o efeito dos mesmos sobre as soluções.

De seguida, na Figura 13, serão apresentados os resultados relativos a esses testes.

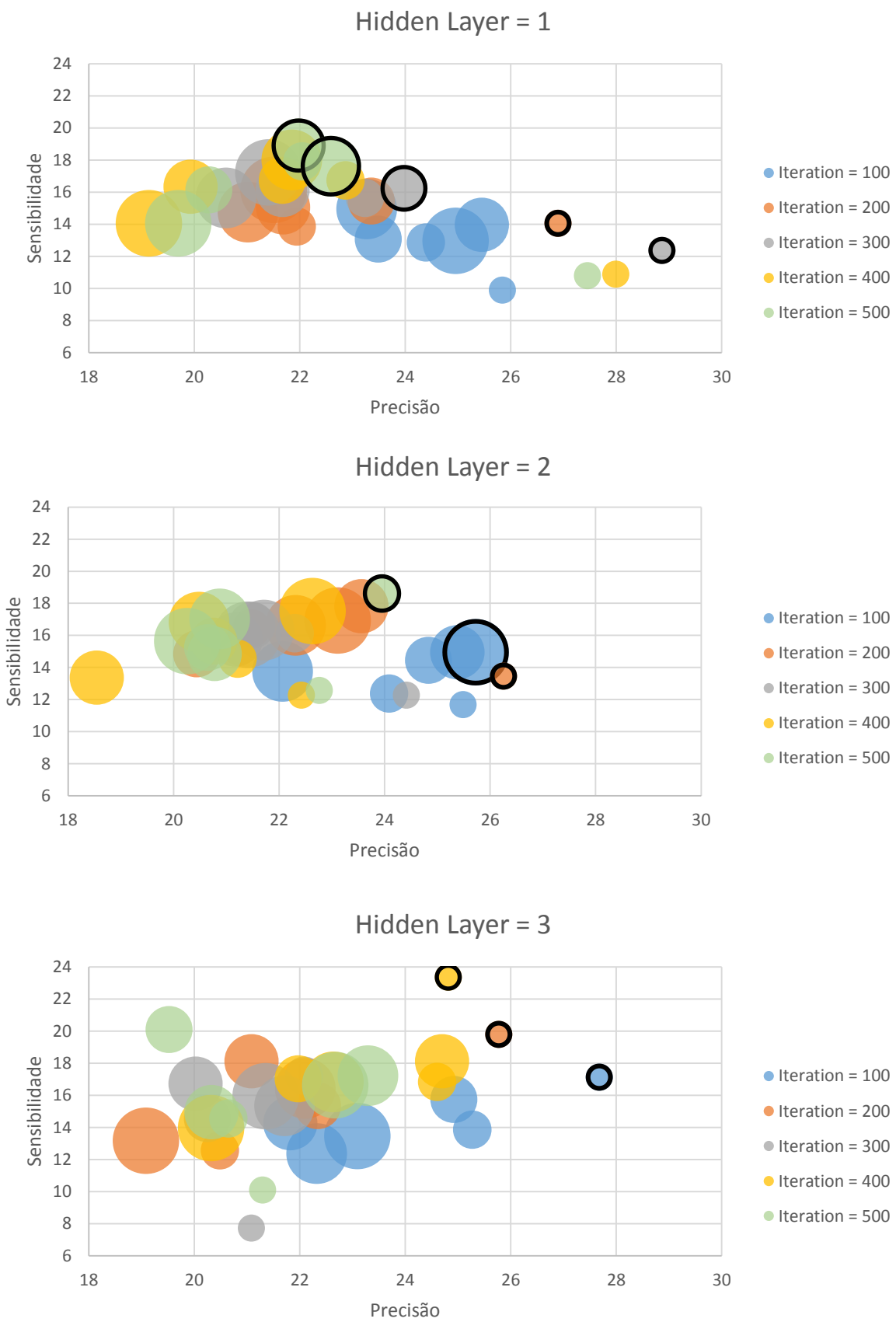


Figura 13 - Resultados do modelo de *Neural Networks* Geral

Os três gráficos (Figura 13) apresentados, representam os resultados obtidos de um conjunto de testes do algoritmo MLP por alteração iterativa de três variáveis seguintes:

- Número de iterações – 100 a 500 em passo de 100;
- Número de neurónios – 5 a 30 em passo de 5;
- Número de camadas intermédias – 1 a 3.

Cada círculo representa uma solução diferente. A posição do seu centro no gráfico representa a sensibilidade e precisão correspondentes a essa solução.

Cada uma das tabelas foi construída com resultados que possuem o número de camadas intermédias referido em cada gráfico. O número de iterações encontra-se representado pela cor dos resultados e, finalmente, o raio do círculo representa o número de neurónios usados – quanto maior, mais neurónios.

Pela análise dos resultados conclui-se que o parâmetro mais difícil de maximizar é a sensibilidade. As melhores soluções estarão sempre localizadas na fronteira superior direita do gráfico, onde se localizam as soluções com a melhor combinação de precisão e sensibilidade, assinaladas com um contorno negro. Entre estas haverá um *trade-off* entre os dois parâmetros.

É interessante assinalar que as melhores soluções em cada caso não são necessariamente as que recorrem a mais neurónios nem iterações, havendo, como no terceiro gráfico, um conjunto soluções superiores com 5 neurónios, nenhuma das quais recorre ao máximo de iterações. Atribui-se este decréscimo da qualidade dos resultados ao sobreajustamento do modelo aos dados de treino, fenómeno descrito no capítulo 2.4.4.

Para definir a melhor solução do conjunto, recorreu-se à métrica F_1 *score*, média harmónica entre a Precisão e Sensibilidade. Assim, obteve.se o gráfico seguinte:

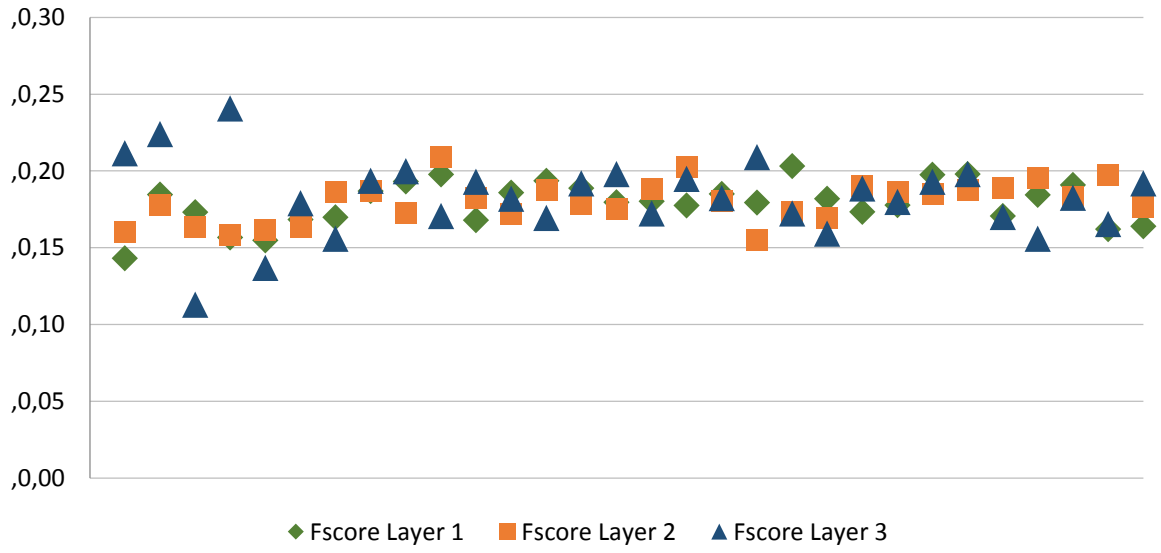


Figura 14 - *F-score* Vs *Hidden Layers*

É de salientar que os dados apresentados na Figura 14 foram obtidos por forma a ter, em cada coluna de valores, o mesmo número de neurónios e iterações. Dessa forma, conclui-se que os resultados com 3 camadas intermédias são, em geral, superiores aos de número inferior de camadas.

A melhor combinação de parâmetros para o modelo MLP geral será apresentada na Tabela 6:

Tabela 6 - Melhores resultados MLP geral

Iterações	Neurónios	Camadas Intermédias	Precisão	Sensibilidade	Taxa de Acerto	F-score
400	5	3	24.82%	23.37%	86.36%	0.24

4.3 Análise de Diagnósticos Secundários e Administrações – Estudo de caso: Pneumonias

Com vista a melhorar os resultados obtidos no modelo geral, como descrito no capítulo 3.4.2, foi analisada, separadamente, uma população composta por episódios cujo diagnóstico principal atribuído é de pneumonia, o que resultou num conjunto de 1626 episódios, com uma taxa de reinternamento de 18.08%. Relativamente a estes, foram analisados os diagnósticos secundários, em ICD-9 e os princípios ativos dos medicamentos administrados, por forma a adicionar campos relevantes ao modelo, sem aumentar demasiado a dimensão do *dataset.*

Apresentam-se, de seguida, as tabelas (Tabelas 7 e 8) com os campos obtidos como resultado dessa análise.

Tabela 7 - Diagnósticos secundários adicionados

Diagnóstico ICD 9	% Reinter. COM Patologia	% Reinter SEM Patologia	Diferenca % Reinternamento Patologia
INFECCAO POR ESTAFILOCOCOS AUREUS RESISTENTE A METICILINA	45.45%	17.51%	27.94%
DIABETES MELLITUS SECUNDARIA S/MENCAO COMPLIC, N/ESPEC.COMO N/CONTR OU N/ESPECIF	44.44%	17.40%	27.04%
DOENCA DE ALZHEIMER	41.67%	16.99%	24.68%
PROLAPSO UTERINO SEM MENCAO DE PROLAPSO DA PAREDE VAGINAL	41.67%	17.52%	24.15%
HEMORRAGIA COMPLICANDO UM PROCEDIMENTO	41.18%	17.45%	23.72%
HISTORIA PESSOAL DE ALERGIA NCOP (EXCEPTO A MEDICAMENTOS)	40.54%	17.18%	23.36%
PSORIASES NCOP	39.68%	16.84%	22.84%
HISTORIA PESSOAL DE TUMOR MALIGNO DA BEXIGA	40.00%	17.49%	22.51%
SINDROMO DE DEPENDENCIA DO ALCOOL SOE	40.00%	17.56%	22.44%
DOENCA HIPERTENSIVA CARDIO-RENAL, N/ESPECIF S/INSUF.CARD C/DRC EST.I-IV OU N/ESPECIFICADA	37.31%	15.04%	22.27%
DOENCAS DA VALVULA TRICUSPIDE	38.46%	17.53%	20.93%
BLOQUEIO DO RAMO ESQUERDO, NCOP	38.46%	17.53%	20.93%
ANGINA DE PEITO NAO CLASSIFICADA EM OUTRA PARTE OU NAO ESPECIFICADA	37.84%	17.24%	20.60%
SINDROMO DE DEPENDENCIA DO ALCOOL, EM REMISSAO	37.50%	17.50%	20.00%
DIABETES MELLITUS C/MANIF.RENAIS, TIPO II OU N/ESPECIF., N/ESPEC.COMO N/CONTROL.	36.36%	17.45%	18.92%
DOENCA RENAL CRONICA TERMINAL	35.71%	17.39%	18.33%
LINFADENITE AGUDA	32.58%	15.94%	16.65%
INFECCAO POR PROTEUS (MIRABILIS) (MORGANII) LNE	33.33%	17.52%	15.81%
INFECCAO POR ESTAFILOCOCOS N/ESPECIF., EM COND.CLASS.OUTRA PARTE, LOCAL N/ESPEC.	33.33%	17.55%	15.78%
DESIDRATACAO	33.33%	17.55%	15.78%
INSONIA, NAO ESPECIFICADA	33.33%	17.58%	15.75%
NEOPLASIA MALIGNA DO RIM, EXCEPTO BACINETE	32.09%	16.45%	15.64%

Foram adicionados ao modelo de previsão de pneumonias os 20 campos apresentados, em coluna, contendo uma indicação binária de 1 ou 0 consoante exista (ou não) um diagnóstico do tipo correspondente associado ao episódio.

A presença dos diagnósticos apresentados na população resulta num acréscimo significativo de risco de reinternamento, pelo que deverá contribuir para a melhoria das previsões.

É de realçar a prevalência de diagnósticos associados a problemas cardíacos e renais (seis dos 20 diagnósticos).

Tabela 8 - Princípios ativos adicionados ao modelo

Principio Ativo	%Reinter. TOMOU Principio	%Reinter NÃO TOMOU Principio	Diferenca % Reinternamento
Dapsona	50.00%	17.64%	32.36%
Ác Cítrico+Óxido De Magnésio+Picossulfato De Sódio	50.00%	17.80%	32.20%
Macronutriente	50.00%	17.84%	32.16%
Metolazona	50.00%	17.88%	32.12%
Bicarbonato De Sódio	43.75%	17.83%	25.92%
Darunavir	41.67%	17.91%	23.76%
Cefoxitina	40.74%	17.70%	23.04%
Amitriptilina	40.00%	17.88%	22.12%
Oxitocina	40.00%	17.95%	22.05%
Aminoác.S+Electrólitos+Glucose+Lípidos	38.89%	17.85%	21.04%
Quetiapina	38.10%	17.82%	20.28%
Ranitidina	36.26%	17.00%	19.26%
Ramipril	34.32%	15.32%	19.00%
Ciproterona	35.19%	17.49%	17.69%
Pantoprazol	34.62%	17.81%	16.80%
Hidroxizina	34.48%	17.78%	16.70%
Morfina	34.38%	17.75%	16.62%
Dinitrato De Isossorbida	33.33%	17.91%	15.42%
Fluoxetina	33.33%	17.94%	15.39%
Memantina	33.33%	17.97%	15.37%

Os campos apresentados foram adicionados ao modelo de forma análoga aos diagnósticos secundários.

Cada um dos princípios ativos administrados encontra-se presente, em média, em 24 episódios da população. Apesar do número reduzido de episódios em que cada um figura, estes têm um efeito significativo relativamente à probabilidade de reinternamento do doente.

É de salientar a prevalência de princípios para o tratamento de condições psiquiátricas (4 princípios) e cardíacas (3 princípios) do conjunto.

4.4 Resultados do Modelo de Pneumonias

Apresentam-se, de seguida, na Figura 15, os resultados relativos ao modelo de previsão de reinternamentos em casos com diagnóstico principal de pneumonia:

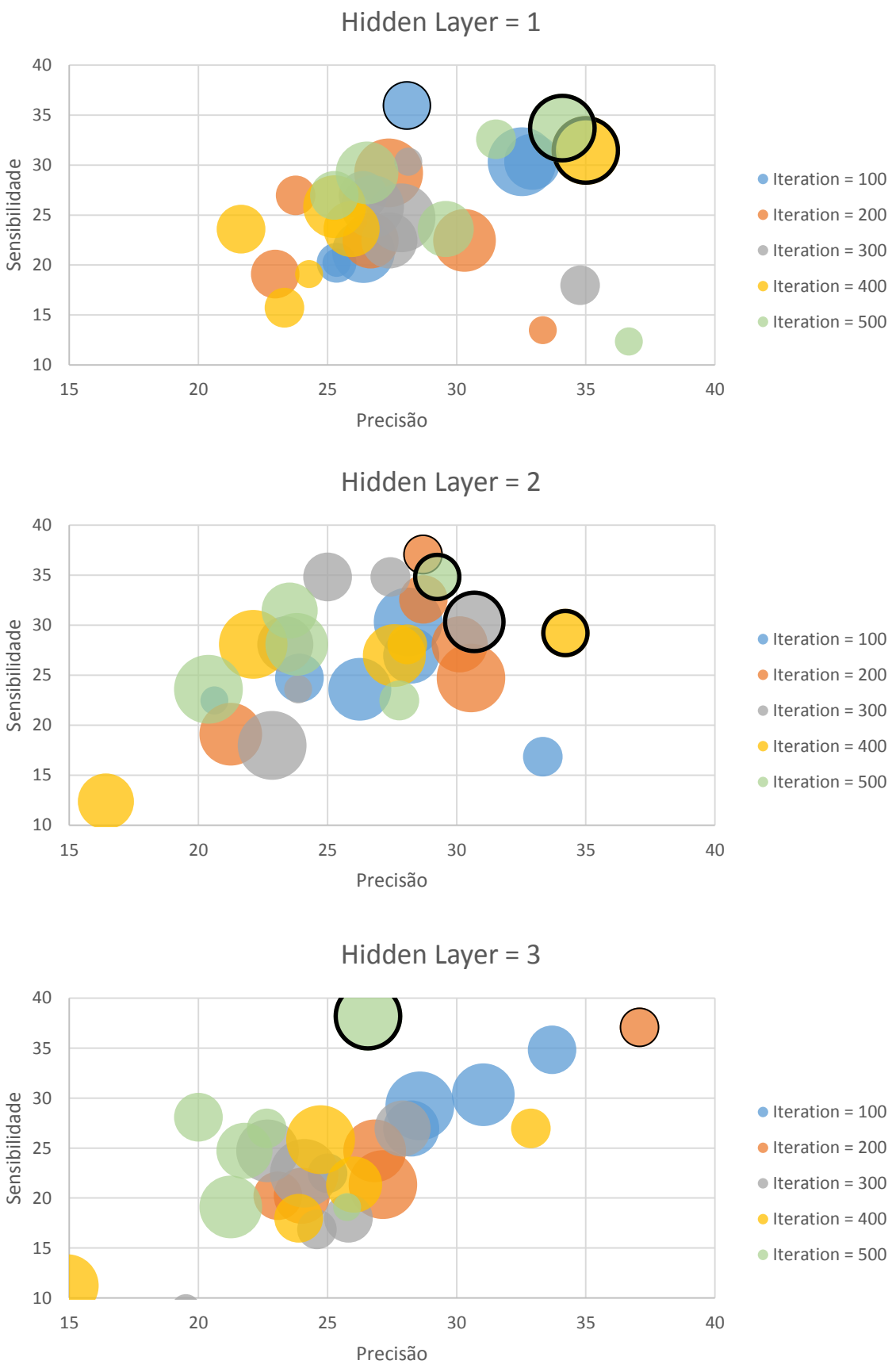


Figura 15 - Resultados do modelo *Neural Networks* Pneumonia

Os três gráficos (Figura 15) apresentam os resultados obtidos de um conjunto de testes do algoritmo MLP, cuja informação representada segue os mesmos moldes dos resultados do modelo geral:

- Número de iterações – 100 a 500 em passo de 100;
- Número de neurónios – 5 a 30 em passo de 5;
- Número de camadas intermédias – 1 a 3.

As cores e raio dos círculos possuem o mesmo significado que anteriormente, bem como os casos assinalados, por forma a facilitar a comparação dos resultados.

Uma vez mais, as melhores soluções em cada caso não são necessariamente as que recorrem a mais neurónios nem iterações, devido aos efeitos do sobreajustamento do modelo aos dados de treino, fenómeno descrito no capítulo 2.4.4.

Conclui-se que, dada a amostra bastante mais reduzida para dados de treino e teste, por vezes, para um número reduzido de neurónios e iterações, o classificador fica viciado e classifica todos os casos como negativos. No entanto, para números de neurónios e iterações superiores o classificador apresenta resultados superiores aos do modelo geral.

Recorreu-se, como anteriormente à métrica F_1 *score* para determinar a melhor solução. Apresenta-se o gráfico (Figura 16) correspondente de seguida:

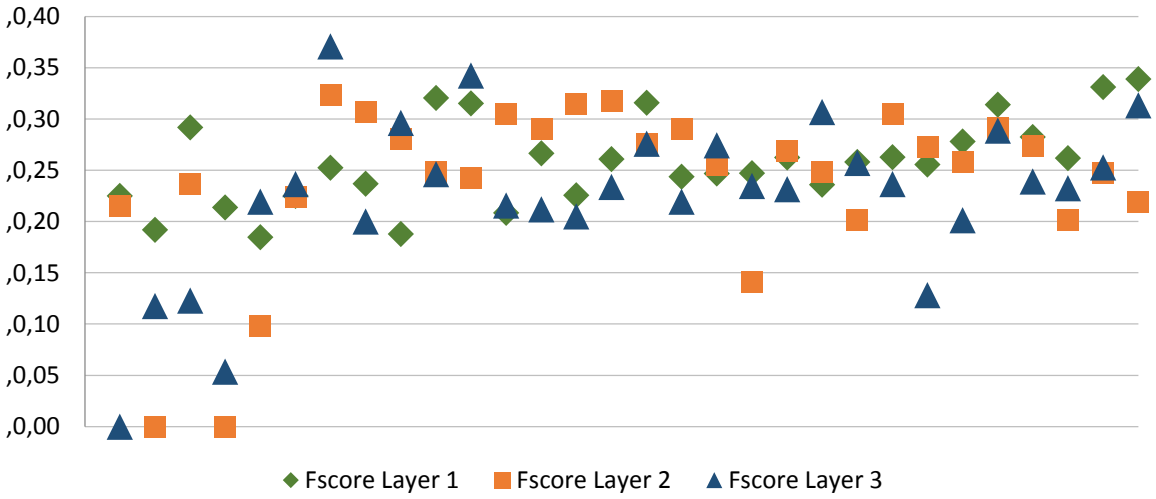


Figura 16 - *F-score* Vs *Hidden Layers*

A melhor solução obtida é apresentada na tabela seguinte (Tabela 9):

Tabela 9 - Melhor resultado do modelo MLP Pneumonia

Iterações	Neurónios	Camadas Intermédias	Precisão	Sensibilidade	Taxa de Acerto	Fscore
200	10	3	37.08%	37.08%	77.10%	0.37

A solução conseguida foi significativamente superior à melhor solução do modelo geral, pelo que se conclui que a abordagem customizada para prever os reinternamentos relativos a uma patologia específica resulta em melhores previsões.

5 Conclusões e perspetivas de trabalho futuro

Ao longo do desenvolvimento deste projeto, foi testada extensivamente a possibilidade de prever a ocorrência de reinternamentos com base nos dados clínicos e administrativos de um doente.

Por comparação com os resultados obtidos noutros trabalhos semelhantes (Tabela 10) (Zolfaghar et al. 2013), os resultados do projeto foram satisfatórios, relativamente ao modelo geral, tendo os resultados do modelo customizado sido ainda melhores.

Tabela 10 - Resultados modelo *Random Forest*

Precisão	Sensibilidade	Taxa de Acerto	F-score
40.47%	1.48%	77.90%	3.01%

Fonte: (Zolfaghar et al. 2013)

Ao longo do trabalho, houve, no entanto, uma série de dificuldades que necessitaram de ser superadas por forma a dar resposta ao problema:

- Muitas vezes foi difícil obter partes da informação presente no *dataset* final, dada esta estar dispersa por vários sistemas de informação. Por essa razão, com cada adição de nova informação, esta necessitou de ser cruzada com o *dataset* já construído e, por vezes, os dados vinham muito incompletos ou incorretos, o que obrigava à identificação e resolução da origem do problema e subsequente nova extração de dados por parte da equipa responsável;
- Muitos campos, apesar de existirem e estarem devidamente tipificados na base de dados, não se encontravam preenchidos, o que levou à tentativa de obter a informação relevante indiretamente, através de outros campos, ou, nos casos em que não foi possível ou seria simplesmente demasiado moroso, o descarte dos campos.

5.1 Trabalho futuro

Sendo o projeto "Previsão de Risco Clínico" um projeto de KDD, este encontra-se sempre sujeito a melhorias iterativas, havendo ainda várias possibilidades de melhoria, que, do ponto de vista dos dados usados para construir o modelo, quer da abordagem utilizada, quer do modelo em si.

Por forma a conseguir melhorar os resultados das previsões obtidas, propõem-se um conjunto de melhorias que poderão vir a ser implementadas:

Implementação de modelos segmentados por idade

Tanto recém-nascidos como crianças jovens são muito pouco suscetíveis a ser reinternados, tendo um perfil de risco significativamente diferente de adultos e idosos. Idosos, por outro lado, encontram-se na população com maior risco, em termos de idades, de ser reinternada.

É ainda de salientar que, analisando as patologias com maior risco e ocorrências e, termos gerais, tratam-se todas elas de patologias que afetam principalmente adultos e idosos, pelo que faria sentido, desta forma, não tratar a população como um todos e segmentá-la. Desta forma, propõe-se:

- Excluir recém-nascidos e crianças do modelo geral;
- Possivelmente, criar um modelo específico para prever reinternamentos de idosos.

Cálculo de métricas relevantes adicionais

Para além das métricas calculadas e testadas ao longo do projeto, existe ainda um conjunto de campos que, com mais tempo e dados mais detalhados, poderiam ser calculados relativamente às administrações:

- Indicação da quantidade total administrada por princípio ativo;
- Indicação da concentração administrada por princípio ativo;
- Indicação de desvios da concentração administrada por princípio ativo relativamente à Dose Diária Recomendada (DDR) do princípio.

Note-se que a DDR é uma fórmula dependente das características do doente e do princípio em questão. Apesar de ser complicado de implementar esta métrica, deverá ser, de entre as métricas referidas, a melhor indicação, relativamente às administrações, da gravidade da condição de um doente.

Análise do histórico de cada doente

Outra adição valiosa para o modelo seria uma análise do histórico de cada paciente, relativamente a outras admissões (não apenas internamentos) no hospital. Tal teria como objetivo a identificação de predisposições e patologias crónicas do paciente.

Poderia ser realizado um estudo relativamente às patologias/predisposições que mais influência têm sobre os reinternamentos e ser introduzido um conjunto de novas variáveis booleanas no modelo por forma a representar essa informação para cada episódio.

Esta análise, no entanto, seria limitada apenas a admissões de doentes no hospital em análise, sendo que, para obter os melhores resultados possíveis, seria necessário ter acesso a todo o historial clínico do paciente, o que não é, de momento, possível.

Obtenção de outros dados relevantes

O trabalho realizado teve por base os dados disponibilizados pelo hospital analisado. Estes dados, no entanto, não contêm toda a informação relevante, ou mesmo informação suficiente para se poder inferir com precisão o estado de cada doente.

Dados tais como tabagismo, a condição financeira e condições de habitação dos pacientes, se têm acesso a medicamentos, bem como outros dados relevantes, de momento, não são registados de nenhuma forma no sistema, apesar de a sua relevância na previsão de reinternamentos ter sido já comprovada (Benbassat and Taragin 2000).

Para além dos campos referidos, não existe também informação tipificada do estado de recuperação da doença (inicial, intermédia, terminal,). Esta, de momento, é registada somente em papel ou em campos de texto livre.

Outro fator importante a adicionar seria as análises (tais como hemogramas) realizadas durante o internamento. Sendo análises que resultem num *output* qualitativo, ou quantitativo para o qual haja valores de referência, desde que devidamente tipificados seriam certamente adições valiosas para o modelo.

Teste de outros algoritmos

Outros algoritmos ainda não testados, bem como outros algoritmos de redes neuronais poderão conseguir resultados superiores aos já testados. Para tal, propõe-se uma revisão de métodos preditivos recentes que possam ter interesse em ser testados.

Para além disso, mesmo para os algoritmos já testados, podem ainda realizar-se testes computacionalmente mais exigentes, sendo que, pela duração do desenvolvimento do projeto, não foram realizados testes com tempo de processamento superior a um dia.

5.2 Aplicação do modelo preditivo - Atribuição de altas por árvore de decisão

Por forma a ter uma aplicação prática, de um ponto de vista financeiro, propõe-se a implementação, no futuro de modelo de decisão teórico para a atribuição de altas.

Para a conceção deste modelo, assuma-se que existe um classificador de casos de reinternamento que possua uma precisão elevada, cujo *output* seja percentual, ao invés de simplesmente binário. Assuma-se também que exista uma fórmula, baseada na análise do custeio do internamento hospitalar que, com base na condição do doente, realize uma estimativa do quanto este custa, por dia de internamento, ao hospital.

Assuma-se ainda, finalmente, a existência de uma forma de cálculo do custo que tem, para o hospital, um episódio semelhante de internamento, para a mesma patologia, de um doente atualmente internado, a quem se está a ponderar dar alta por indicação médica. Por falta de uma fórmula desenvolvida para tal, assuma-se que o custo será igual ao indicado na tabela de preços por GDH, do documento relativo a contratos-programa da ACSS (ACSS (Ministério da Saúde) 2015).

Assumindo a existência de todos os valores referidos, seria possível construir uma árvore de decisão (Figura 17) que pesasse, aquando da alta, o valor esperado, em gastos, para o hospital, de manter o doente mais um dia contra o valor esperado de lhe atribuir alta.

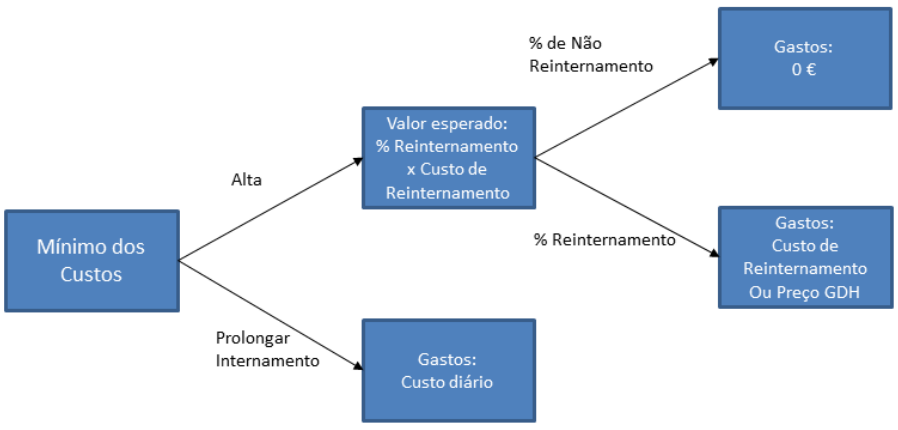


Figura 17 - Árvore de decisão de prolongamento de altas

Este processo, caso calculasse que o valor esperado de manter o doente mais um é superior (resulta em menos perdas) ao de atribuir uma alta imediatamente (pelo risco de reinternamento associado), iria automaticamente atribuir mais um dia, por prevenção, para o internamento do doente. Deixaria de atribuir mais dias quando o risco de reinternamento fosse baixo o suficiente tal que o valor esperado de prejuízo pelo reinternamento fosse mais baixo que o de manter o paciente mais um dia.

Dessa forma, estaria a poupar-se fundos ao hospital e, simultaneamente, providenciar melhor cuidado médico a casos com maior risco de reinternamento.

5.3 Considerações finais

A aplicabilidade de KDD a esta área é um tópico relativamente recente, mas que tem ainda muito para explorar, como se tem vindo a constatar pelo trabalho desenvolvido ao longo dos últimos anos, tanto em Portugal como no estrangeiro.

Certamente continuará a ser explorada, dado o relevo atual que tem a otimização de recursos hospitalares, tanto pelo seu peso elevado na economia, como pela qualidade dos serviços prestados em hospitais e, possivelmente, trará benefícios a toda a população.

Referências

ACSS (Ministério da Saúde). 2015. "Contrato-Programa 2015." http://www.acss.min-saude.pt/Portals/0/Metodologia_HH_ULS_2015.pdf.

Benbassat, Jochanan, and Mark Taragin. 2000. "Hospital Readmissions as a Measure of Quality of Health Care." *American Medical Association* 160: 1074–81.

Boquinhas, José Miguel. 2012. *Políticas E Sistemas de Saúde*. Edited by Almedina. 1ª edição.

Borges, Cláudia Medeiros. 2011a. "Doentes Equivalentes - Portal Da Codificação Clínica E Dos GDH." http://portalcodgdh.min-saude.pt/index.php/Doentes_equivalentes.

———. 2011b. "Grupos de Diagnósticos Homogéneos (GDH) - Portal Da Codificação Clínica E Dos GDH." http://portalcodgdh.min-saude.pt/index.php/Grupos_de_Diagn%C3%B3sticos_Homog%C3%A9neos_(GDH).

———. 2011c. "Índice de Case-Mix (ICM) - Portal Da Codificação Clínica E Dos GDH." http://portalcodgdh.min-saude.pt/index.php/%C3%8Dndice_de_Case-Mix_(ICM).

Breiman, Leo. 1999. "Random Forests-Random Features."

Campos, Alexandra. 2014. "Portugal É O País Da União Europeia Onde Mais Se Morre Por Pneumonia - PÚBLICO." http://www.publico.pt/sociedade/noticia/portugal-e-o-pais-da-uniao-europeia-onde-mais-se-morre-por-pneumonia-1678845.

Castro, Ricardo Alves De Sousa. 2011. "Benchmarking de Hospitais Portugueses: Modelação Com Data Envelopment Analysis." *FEUP - Faculdade de Engenharia* Dissertaçã: XII, 110 p., 30 cm. http://hdl.handle.net/10216/62130.

Fundação Francisco Manuel dos Santos (Instituição). 2015. "PORDATA - Estado: Execução Orçamental." http://www.pordata.pt/Portugal/Ambiente+de+Consulta/Tabela.

Gama, João, André Carvalho, Katti Faceli, Ana Lorena, and Márcia Oliveira. 2012. *Extração de Conhecimento de Dados - Data Mining*. Edited by Edições Silabo.

Glintt. 2013. "Missão - Glintt.com." http://www.glintt.com/missao.

Gomes, Carlos Godinho. 2014. "Optimizing Operating Room Planning - A Data Mining and Optimization Approach."

Gomes, Lucy. 2001. "Fatores de Risco E Medidas Profiláticas Nas Pneumonias Adquiridas Na Comunidade." *Jornal de Pneumologia* 27 (2). Sociedade Brasileira de Pneumologia e Tisiologia: 97–114. doi:10.1590/S0102-35862001000200008.

HCUP. 2015. "HCUP-US Tools & Software Page." http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.

Helm, Jonathan E., Adel Alaeddini, Jon M. Stauffer, Kurt M. Bretthauer, and Ted A. Skolarus. 2013. "Reducing Hospital Readmissions by Integrating Empirical Prediction with Resource Optimization." doi:10.1053/j.gastro.2012.12.021.This.

Jweinat, Jillian J. 2010. "Hospital Readmissions under the Spotlight." *Journal of Healthcare Management / American College of Healthcare Executives* 55 (4): 252–64.

Kanel, Keith, Susan Elster, and Colleen Vrbin. 2010. "PRHI Readmission Brief - Brief II: Patterns of Hospital Admission and Readmission Among HIV-Positive Patients in Southwestern Pennsylvania," no. December: 1–12.

Kanel, Keith T., Susan Elster, and Colleen Vrbin. 2011. "PRHI Readmission Brief - Chronic Obstructive Pulmonary Disease," no. December: 1–12.

Kohavi, R, and Dan Sommerfield. 1995. "Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology." *First International Conference on Knowledge Discovery and Data Mining*, 192–97. http://www.aaai.org/Papers/KDD/1995/KDD95-049.pdf.

Lopes, Fernando. 2009. "Episódio de Internamento - Portal Da Codificação Clínica E Dos GDH." http://portalcodgdh.min-saude.pt/index.php/Epis%C3%B3dio_de_internamento.

Marbán, Óscar, Gonzalo Mariscal, and Javier Segovia. 2009. "A Data Mining & Knowledge Discovery Process Model." *Data Mining and Knowledge …*, no. February: 1–17. http://cdn.intechopen.com/pdfs/5937/InTech-A_data_mining_amp_knowledge_discovery_process_model.pdf.

OMS. 2005. "Portal Da Saúde - O Que São Doenças Crónicas?" http://www.portaldasaude.pt/portal/conteudos/enciclopedia+da+saude/ministeriosaude/doencas/doencas+cronicas/doencascronicas.htm.

Pereira, Gonçalo. 2014. "Data Mining Na Prevenção de Riscos Clínicos."

Piatetsky, Gregory. 2014. "CRISP-DM, Still the Top Methodology for Analytics, Data Mining, or Data Science Projects." http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

Riedmiller, Martin, and Heinrich Braun. 1993. "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm." In *IEEE INTERNATIONAL CONFERENCE ON NEURAL NETWORKS*.

Silva, Vanessa. 2011. "Reinternamento - Portal Da Codificação Clínica E Dos GDH." http://portalcodgdh.min-saude.pt/index.php/Reinternamento.

Sousa-Pinto, Bernardo, Ana Rita Gomes, Andreia Oliveira, Carlos Ivo, Gustavo Costa, João Ramos, Joel Silva, et al. 2013. "Reinternamentos Hospitalares Em Portugal Na Última Década." *Acta Medica Portuguesa* 26 (6): 711–20.

Sperandio, Fabrício Reuter. 2015. "Large Scale Elective Surgery Scheduling under Uncertainty."

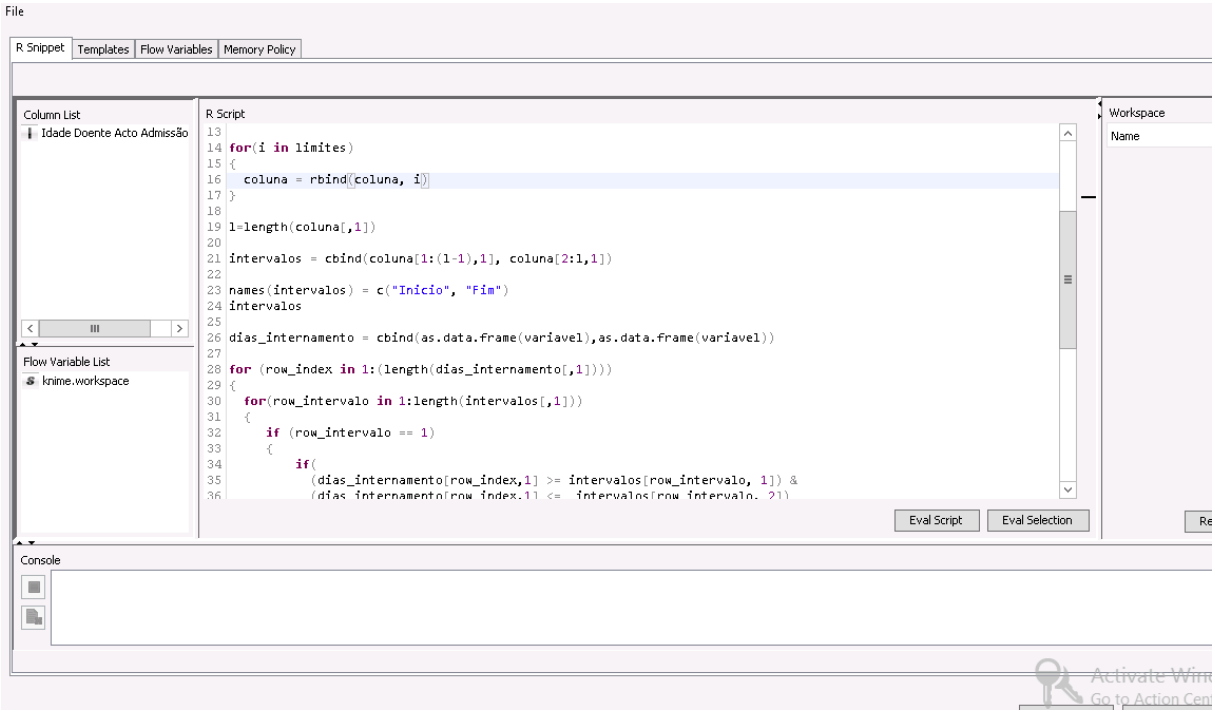
Técnico, Núcleo de Engenharia Biomédica do Instituto Superior. 2008. "No Title." http://nebm.ist.utl.pt/entidades/ent/4.

Zheng, Bichen, Jinghe Zhang, Sang Won Yoon, Sarah S. Lam, Mohammad Khasawneh, and Srikanth Poranki. 2015. "Predictive Modeling of Hospital Readmissions Using Metaheuristics and Data Mining." *Expert Systems with Applications* 42 (20). Elsevier Ltd: 7110–20. doi:10.1016/j.eswa.2015.04.066.

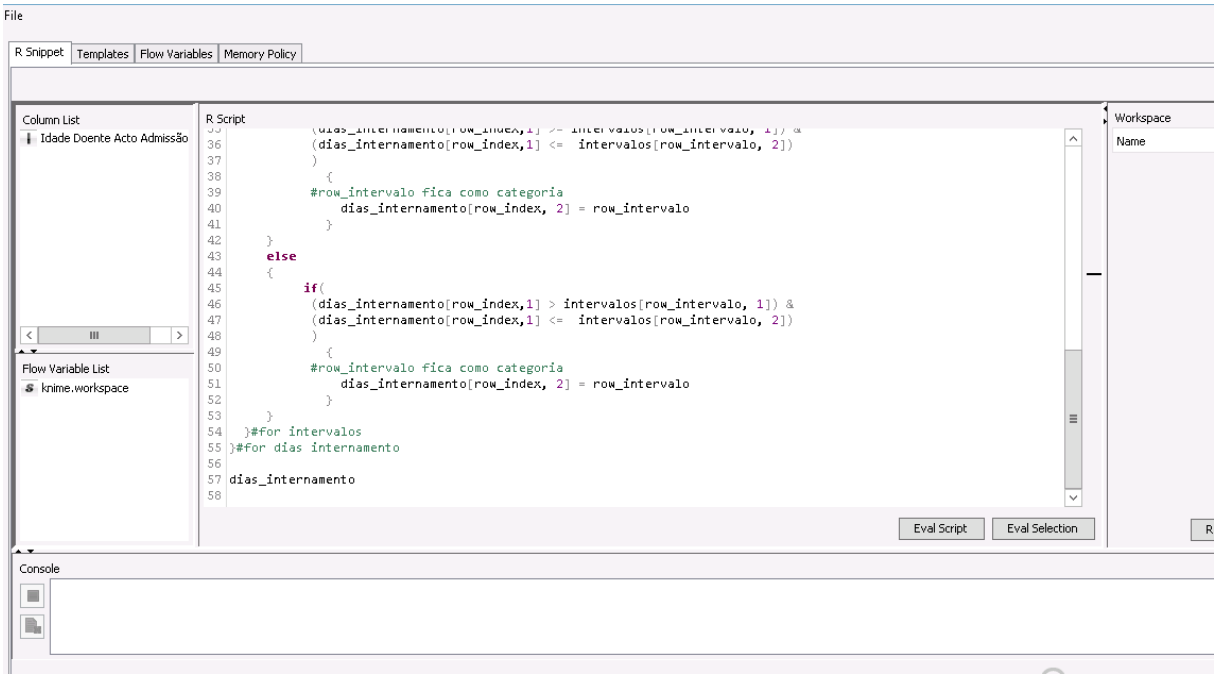
Zolfaghar, Kiyana, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Si Chi Chin, and Brian Muckian. 2013. "Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients." *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, 64–71. doi:10.1109/BigData.2013.6691760.

ANEXO A: Código elaborado em R

- Exemplo 1:

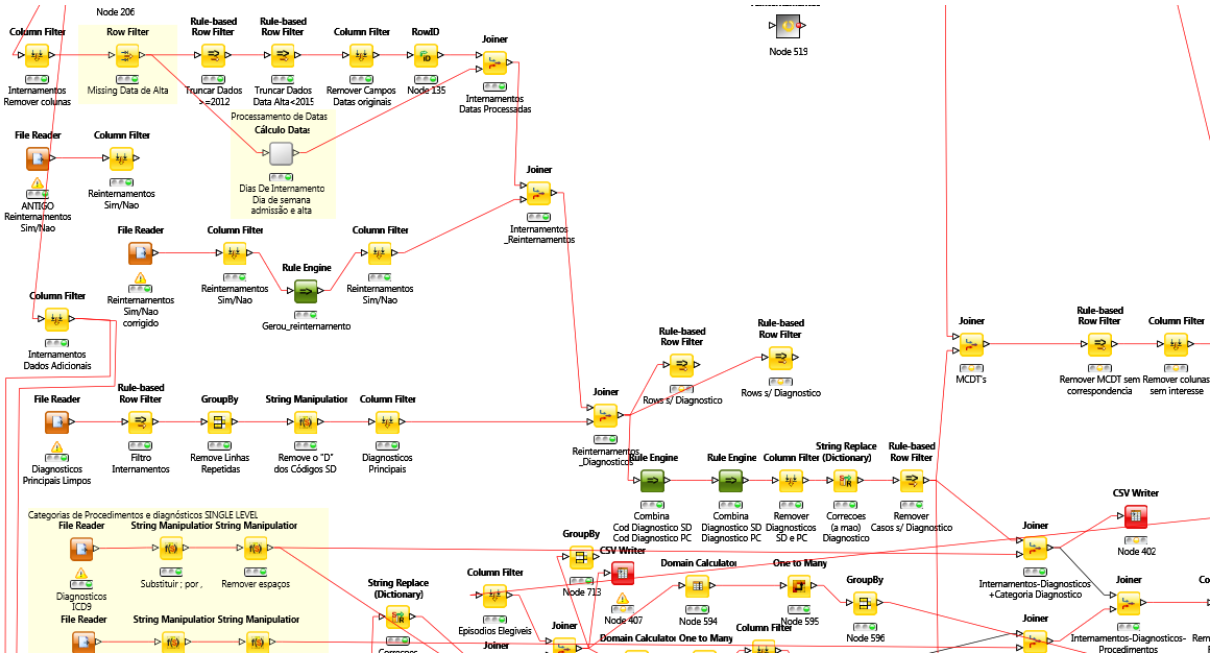


- Exemplo 2:

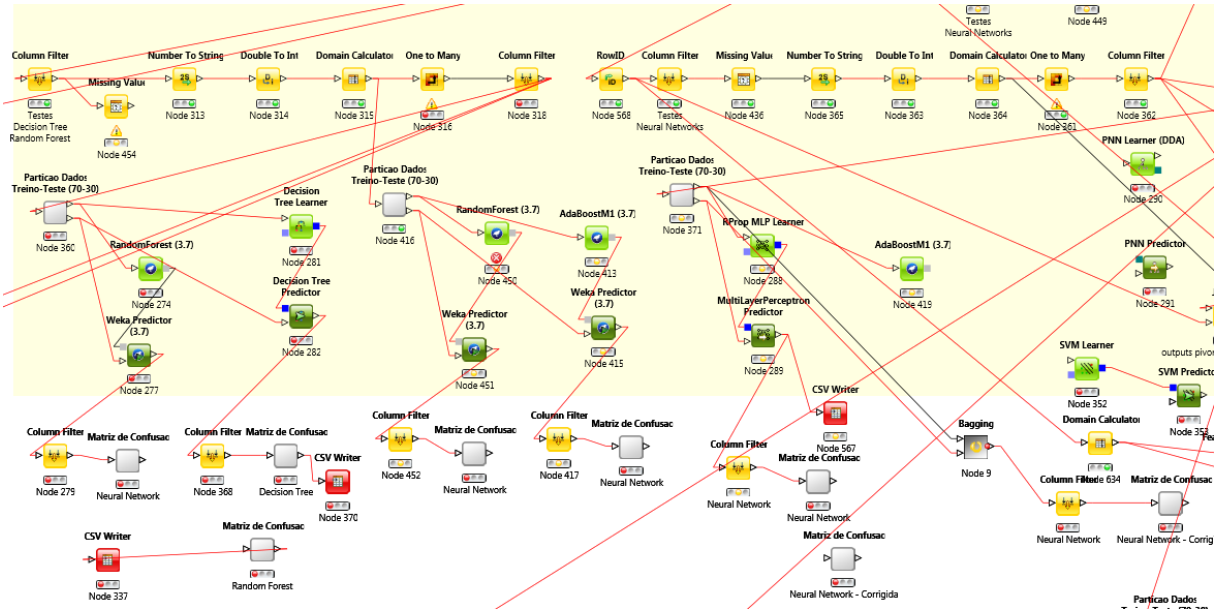


ANEXO B: Diagramas do Projeto em Knime

- Limpeza de dados e junção de tabelas:



- Modelo geral:



- Modelo Pneumonias: